



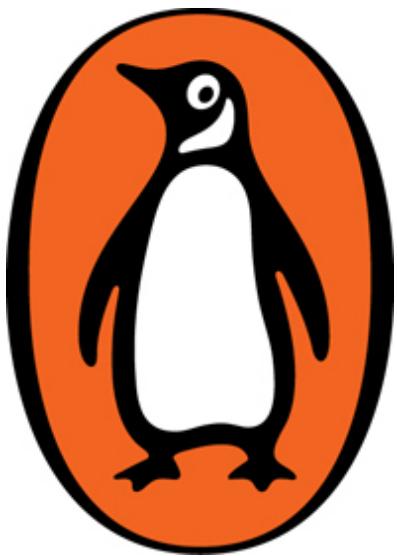
From the bestselling author of
Winning in the Digital Age

MASTERING THE DATA PARADOX

THE KEY TO WINNING
IN THE AI AGE

NITIN SETH

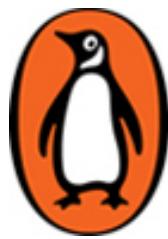
Co-founder and CEO, Incedo Inc.



NITIN SETH

MASTERING THE DATA PARADOX

The Key to Winning in the AI Age



PENGUIN BOOKS

Contents

Introduction: AI Age and the Data-First World

SECTION I

UNDERSTANDING THE DATA-FIRST WORLD

1. Data Explosion: An Unprecedented Phenomenon
2. Data, the Fuel for the Digital Age: Impacting Every Aspect of Life and Business
3. Value Reimagined: Framework for Realizing Transformational Value from Data
4. The Data Paradox: Deluge and Drought
5. The Root Cause: More Logical than Physical

SECTION II

MAXIMIZING VALUE IN THE DATA-FIRST WORLD

6. A Unified Solution Framework: Thirteen Mantras for Data Success
7. Define the Business Problems: Trees, Trees, Trees!
8. Multi-Source Data: It's Not a Lake, It's an Ocean
9. Real-Time Data: Action at the Speed of Light
10. Proprietary Data: The Secret Sauce for Competitive Advantage
11. Modern Data Stack: Scalable Platforms, Front-to-Back Integration
12. Data Quality: It's More than What Meets the Eye
13. Data Products: The Elusive Link between Data, Action and Impact
14. Agility: Mastering Two-Speed Execution

- [15. Data Democratization: Empowering with Data](#)
- [16. Data Security: Biggest Threat to the Data Revolution](#)
- [17. Organizational Alignment: Who Owns the Data?](#)
- [18. Data Culture: Beware of the HiPPO Effect!](#)
- [19. Data Talent: Talent Continues to Be Key in the AI Age](#)

SECTION III

DATA FOR INDIVIDUALS AND BEYOND

- [20. Moving from Enterprises to Individuals and Society](#)
- [21. The World of Hyper-Personalization: Segment of One](#)
- [22. Data for Better Decision-Making: Live Better with Data](#)
- [23. Information and Wisdom: Reflect and Recognize Patterns](#)
- [24. Data Sharing vs Data Privacy: Find the Right Balance](#)
- [25. Digital Engagement vs Mental Health: Connect with Your Inner Self](#)
- [26. Data Collaboration for a Better World: A New Vision for Global Collaboration](#)
- [27. Data as a Source of National Competitive Advantage: Twenty-First-Century National Asset](#)

In Conclusion: Mastering the Data Paradoxes to Win in the AI Age

Acknowledgements

Glossary of Data Concepts

Footnotes

- [1. Data Explosion: An Unprecedented Phenomenon](#)
- [4. The Data Paradox: Deluge and Drought](#)
- [9. Real-Time Data: Action at the Speed of Light](#)
- [11. Modern Data Stack: Scalable Platforms, Front-to-Back Integration](#)
- [14. Agility: Mastering Two-Speed Execution](#)
- [15. Data Democratization: Empowering with Data](#)
- [25. Digital Engagement vs Mental Health: Connect with Your Inner Self](#)

26. Data Collaboration for a Better World: A New Vision for Global Collaboration

27. Data as a Source of National Competitive Advantage: Twenty-First-Century National Asset

In Conclusion: Mastering the Data Paradoxes to Win in the AI Age

Notes

Follow Penguin

Copyright

ADVANCE PRAISE FOR THE BOOK

'As a hedge fund manager/investor, information has been the lifeblood of my business career. Extracting the value from data in myriad different situations is the essence of beating the market. [Nitin] Seth's *Mastering the Data Paradox* is loaded with common sense, wisdom and detailed expertise on all aspects of data. As we enter the age of artificial intelligence, every professional needs to understand this crucial subject'—**David Cohen, founder and owner, Simcah Management**

'After exploring the complex technology trends that shape the business world in *Winning in the Digital Age*, Nitin Seth now offers a broad perspective on the fuel that powers the digital age and is also, paradoxically, its output—data ... in all sizes and shapes, the generation and usage of which come with a variety of imperatives and implications, especially as it applies to generative AI. *Mastering the Data Paradox* explores these with significant depth but in the simple language that Nitin has mastered by leveraging his mixed background as a consultant and a practitioner. Given Belden's focus on data engineering, I found the book instructive in terms of better understanding how our customers' needs are evolving'—**Ashish Chand, president and CEO, Belden Inc.**

'Data, which has become synonymous with digital transformation that spans everything from corporate strategy to personal lives, has not always delivered on the promise of making everything more productive and experiences better. This book directly addresses the vast opportunity represented by data and a structured approach to translating this into outcomes at all levels. Nitin Seth, with deep professional experience in this field, is able to translate abstract concepts into practical examples and provide an objective and logical approach throughout the book that will help readers from all backgrounds get more informed and draw out the parts most relevant to their situation. The practical frameworks in the book

show the interconnectedness of people, processes and culture, which is ultimately key to translating the potential from data proliferation into insights and outcomes for companies as much as individuals. In a fast-evolving space with the advent of artificial intelligence, the book provides a timely and comprehensive guide to understanding data and, more importantly, using it to achieve your objectives’—**Dhrupad Trivedi, CEO, A10 Networks**

‘The explosion of data, combined with rapidly advancing artificial intelligence, is the biggest challenge and opportunity in today’s business world—you can ride this powerful wave or let it wash over you. In *Mastering the Data Paradox*, Nitin provides an intelligent, specific approach to mastering this convergence and creating new value in the emerging AI Age’—**Michael J. Ragunas, chief technology officer, Cetera Financial Group**

‘Mastery of data is the lifeblood of any organization, regardless of industry vertical; and ironically, it is one of the most significant yet fundamental threats to those who do not find a structured approach to rein in its complexity. Nitin Seth’s first book, the bestselling *Winning in the Digital Age*, is now enhanced by the context of his personal passion and practitioners’ experiences in understanding the overlap of the strategic complexity of data and the value of a sound perspective to drive success in a digital world. Sometimes, of course, timing is everything, and with the proliferation of advanced data tooling and AI-driven disruption, *Mastering the Data Paradox* provides a foundation for understanding the historical context of the data explosion, and how to manage what Seth refers to as the “dark side of data” to deliver true value. As a technologist and a mathematics major with a passion for data science and its practical use to drive results, I was pleased to see the historical technical context and humbled to learn from all of the vast experiences and practical advice Seth presents to the reader. Effectively, this is a guidebook to navigate the truism that “Big Data is only valuable if it leads to Big Decisions”, whether to enhance your work environment; or uniquely, how data can be leveraged to advance “digital detox” to

improve your personal life!'—**Mukesh Mehta, EVP and chief information officer, AssetMark**

'Well-structured and easy to read, this handbook is a distillation of experience and research, addressing one of the important issues of our time. Nitin Seth elegantly diagnoses the challenges of synthesizing the vast amount of data now available and provides effective solutions to harnessing this power as AI becomes integral to business and society. The focus on logic to shape the right questions is especially helpful. Practitioners, entrepreneurs and administrators will find these lessons invaluable'—**Leo Puri, senior financial services executive, adviser and investor**

'A highly recommended read for those seeking to master the art of navigating the paradoxical world of data. Nitin has written an expansive book explaining the opportunities and complexities of the world of data and its implications on individuals and businesses, and in shaping nations. The book breaks the trade-off by explaining a highly complex topic in an accessible, easy-to-understand manner'—**Dinesh Khanna, managing director and senior partner, Boston Consulting Group**

'Nitin Seth carries on from his first book *Winning in the Digital Age* to a much-needed playbook on how to think about data, the oil of this digital age. His practical experience in using data to run companies and his strong theoretical understanding of the new data ecosystems underpinning the AI revolution guide the reader through the complex journey of finding sufficient amounts of the right data for the desired purpose. The paradox of Big Data, where more is not always better and often worse, is clearly resolved and will surely help many people lost in the data maze'—**Anurag Agrawal, dean, biosciences and health research, Ashoka University**

'The release of ChatGPT in November 2022 unleashed a new technology wave that promises to be as big, if not bigger than the internet itself. Since then, corporate executives have been rushing to

figure out what generative AI could mean for their businesses both in the short- and long-term. Data is at the centre of this new imperative, and Nitin's book takes a deep dive on various opportunities and challenges faced by corporations regarding this vital asset. As in his earlier book on digital transformation, Nitin lays out a clear data framework for enterprises as they embark on an AI journey. The chapters on data security and democratization are of particular importance, as they are central to enterprise AI adoption amid fears of large language models training on proprietary corporate data, and then sharing the results publicly or with a rival'—**Anurag Rana, senior technology analyst, Bloomberg Intelligence**

'Data is the key behind AI and digital transformations. While there is a lot of great work happening in technology, structuring, capturing and managing data is key to the deployment of these new transformational tools. Going beyond generalities around data, this book delves deeper to uncover the nuances and paradoxes of the data world, offering a fresh perspective and exploring often-overlooked nuances. This book resonates well with individuals and practitioners across the spectrum, serving as a valuable resource for those seeking to wrap their heads around the data complexities. Nitin's expertise and unique perspective act as a much-needed guide to navigate the Big Data and AI landscape. Nitin has once again delivered a remarkable piece of work with this latest book, following the success of *Winning in the Digital Age*'—**Som Mittal, former president and chairman, Nasscom**

*To my dear wife, Arpna, and my children, Karmishtha,
Devishi and Pragun, who are the pillars of my life—
for believing in me and inspiring me to do better every day.*

*To my magnificent German Shepherd, Champ, for being
the selfless soul sitting beside me for the many hundred
nights it took me to work on this book!*

Introduction

AI Age and the Data-First World

'AI is one of the most important things humanity is working on. It is more profound than, I dunno, electricity or fire.'

—Sundar Pichai,
CEO, Alphabet and Google

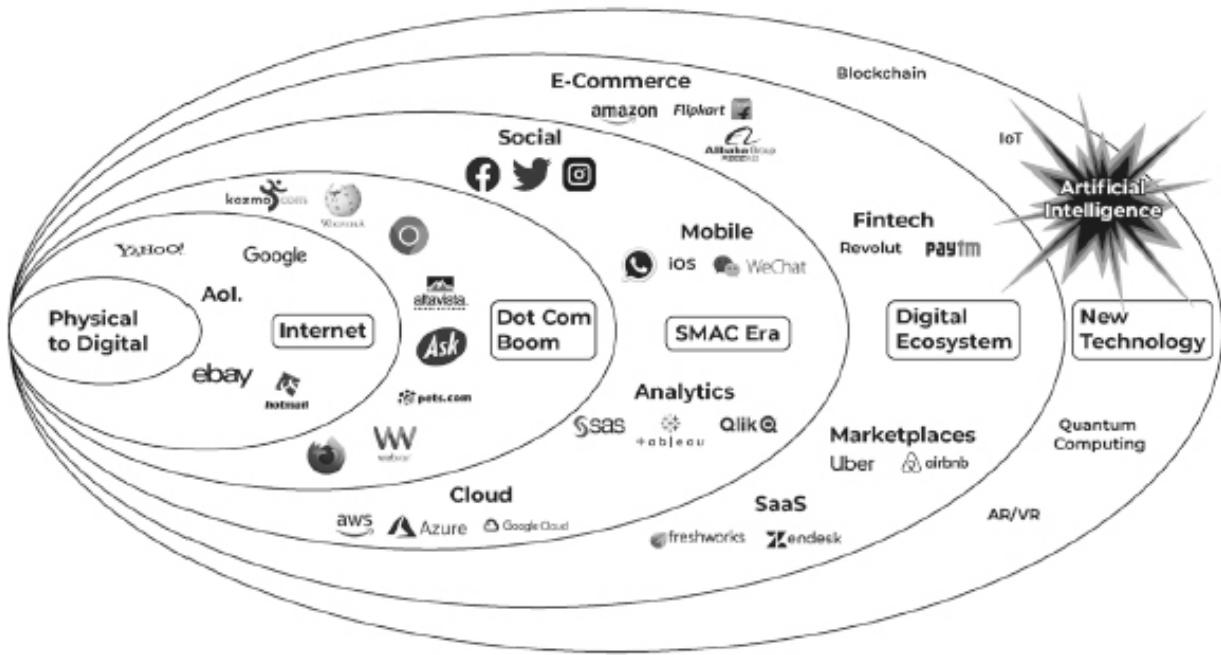
The AI age is here

In my previous book, *Winning in the Digital Age*, I defined the digital age as a wide-ranging set of technology trends that have evolved over the years, impacting every aspect of life and business. For me, the digital age comprises a broader set of events happening, a continuously evolving set of bubbles starting with digitization to the age of the internet, to the dot-com boom, then the SMAC (social, mobile, analytics and cloud) era, the digital ecosystem era and finally the era of new technologies like artificial intelligence (AI), blockchain, the internet of things (IoT), quantum computing, etc. These and many other new technologies have been growing significantly over the past few years. But one key technology—AI—has witnessed disproportionate growth in recent times.

When you look at the infographic below, it becomes evident that the advancements across various domains within the digital age, from digitization to the new technologies era, would not have been possible without the exponential growth of data. So, a significant aspect of this digital age is data, which has brought us to the data-first world where organizations recognize the immense value of data and aim to prioritize it as a core asset. Now, the data-first world has

laid a strong foundation for the AI age by providing abundant and diverse data, further amplified by the advent of generative AI (Gen AI), which literally unlocked the data of the world. And therefore, within the digital age, the data-first world and the AI age are significantly intersecting and reinforcing each other. Digital technologies have led to the explosion of data and indeed the foundation of the AI age. And AI in itself can significantly enhance many aspects of the data-management cycle as well as the digital experience. This interplay between digital, data and AI is truly fascinating and leading to the rapid onset of the AI age.

AI has exploded in the last few years to become one of the key drivers of the Digital Age



Of course, AI is not a new concept. It has been around for more than seventy years, with the concept emerging as early as the 1950s. And over the years significant progress has been made in this field. But with all its promises and possibilities, scaling AI has been a major issue. This is because for AI to be successful, three components are crucial: data, computational power and algorithms. While algorithms and computational power are more solvable, data has remained the most critical challenge, becoming more complex

by the day. An AI model is only as good as the data that powers it. In the enterprise context, despite the availability of Big Data, organizations often encounter challenges in accessing sufficient and relevant data. Although the concept of Big Data suggests abundant data resources, accessing the right data that aligns with specific business needs can be a complex task. While digital natives such as Amazon and Netflix with massive customer behaviour data have an advantage, for most companies across industries, it isn't easy to get the data required, clean it and build AI models on it. This is why, in most cases, AI has not grown beyond proof-of-concept or the experimentation stage to scale, other than a few specific use cases like personalization.

But since the launch of ChatGPT in November 2022, the story has changed. Gen AI has massively unlocked AI capabilities. How? In September 2021, all the data on the internet was downloaded. And by accessing and working with this immense data set, powered by sophisticated neural network models, Gen AI has unlocked a higher level of intelligence and generative capabilities. And because we now have a way to literally leverage the data of the world, the challenge around the availability of data has been significantly addressed. It has led to a significant leap in AI capabilities, moving beyond traditional interpretive capabilities of AI, like pattern recognition, content categorization, etc., to the generative capabilities of Gen AI, like creating new and original content. With that the 'AI age' is now very real and it is happening right now. The genie is out of the bottle!

As Malcolm Gladwell said, 'The tipping point is that magic moment when an idea, trend, or social behaviour crosses a threshold, tips, and spreads like wildfire.' The advent of Gen AI is that tipping point. It has opened new opportunities to leverage the collective wisdom of crowds and tap into the infinite possibilities of data. Gen AI also contributes to the expansion of data sources, by adding new content to the pool of data available for analysis and innovation. Furthermore, Gen AI offers a clear advantage by providing a starting point for problem-solving, reducing the ambiguity of where to begin.

So, the entire trajectory of the AI age hinges on data because it is that underlying element that would help in contextualizing and maximizing the potential of AI. Combining Gen AI models with the vast amount of data within an organization would allow for the creation of specific and actionable insights that are relevant to the context of the business. This integration would help drive both efficiency and innovation within the organization.

Combining Gen AI with individual data would enhance the user experience, making it more relevant for individuals, whether it's personalized product recommendations, customized news feeds or personalized marketing messages. By analysing data on every individual, such as their preferences, behaviours and past interactions, etc., AI can generate tailored content and suggestions that are specifically designed for them. So, while Gen AI models alone may not be enough, contextualizing them by integrating them with enterprise and individual data, as well as existing models and systems, can unlock synergies and allow for the creation of more comprehensive and impactful AI-driven solutions.

I foresee months and years ahead marked by tremendous experimentation and development that would result in significant progress, not just in Gen AI but in other aspects of AI as well. Industry- or function-specific Gen AI models will emerge, meshing with enterprise and individual data, enabling organizations to unleash the true potential of AI.

Hence, this book delves into the entirety of data, uncovering opportunities, challenges and solutions across enterprise, individual and macro levels. This exploration, I believe, is critical to uncover the pivotal role of data, which is the necessary foundation for organizations to thrive in the data-first world and triumph in the AI age.

My journey into the data-first world

The past twenty-five years of my professional and personal life have been a fascinating journey of experiments with data. Often caught in the thick of things, where data played a critical role in business

success, I spearheaded some highly innovative and radical ways to make an impactful change with data. My experiences have led me to a firm conviction that data is one of the greatest opportunities in the AI age and data-first world that we are in, but it equally presents some very difficult challenges. And despite the undeniable importance of data, it is a topic that is not well understood.

This book is a synthesis of my experiments and experiences with data, where I have tried to create a practical playbook that should be helpful to both seasoned executives and young professionals as well as students on how to win in a data-first world. These experiences include building perhaps the most unique and innovative knowledge centre in the world for McKinsey, a top management consulting firm; leading global offshore operations and setting up the strategy and planning and analytics functions at Fidelity International, one of the largest active asset managers in the world; leading Flipkart, an e-commerce major and the biggest start-up in India at the time, at one of its most tumultuous times as a chief operating officer (COO); and driving multiple digital initiatives for a number of Fortune 500 companies as the CEO of Incedo Inc., a US-based consulting, data science and technology services firm, for more than a quarter of a century now. But before I delve deeper into my journey, let me first set the context for why I was convinced I must dedicate an entire book to the topic—DATA.

The current era of the digital revolution, beginning somewhere in the late twentieth century, is marked by the widespread adoption of connected digital technologies that have significantly transformed various aspects of our lives and businesses. The digital economy now contributes to more than 15 per cent of global gross domestic product (GDP), growing 2.5x faster than overall GDP, on average, over the past decade.¹ As my fingers fly across the keyboard, four of the top five publicly traded companies in the world are technology companies, each with a market capitalization of over \$1 trillion.²

Today, we stand at a crucial juncture in the digital revolution happening all around us. Decades of technological progress have democratized access to digital technologies. Mobile phones and

computers have ingrained digital into our daily lives, influencing every aspect of it, whether we like it or not. Historically, every period of human endeavour has had these pivotal moments, defining human progress, from the crafting of fire the very first time to the reusable rockets of SpaceX. But most of us typically overlook these pivotal moments in the pursuit of immediate concerns and endeavours, only to recognize their significance through hindsight and the passage of time.

In my previous book, *Winning in the Digital Age*, I shared my learnings of the past twenty-five years working both as a consultant serving other organizations and senior executive leading business and digital transformation at various organizations. It was an attempt to bring clarity to the complex and vast topic of digital, by sharing practical insights and best practices on the various aspects of digital transformation.

During this journey, the most prominent aspect of the digital revolution that emerged for me was data. I have been a wide-eyed witness to the phenomenal growth of data and the unprecedented opportunities that it brought with it. I have seen and, in fact, been one of the early voices in propagating the tremendous opportunities that data and analytics could bring to an organization. I have seen how data and analytics have transformed the world of consulting and beyond and have been blown away by the sheer force that data has become today.

But every story has two sides—I have also been a teary-eyed witness to the havoc that the inability to manage data well can cause in an organization. As data grew in size, shape and form, it became a relentless force in itself. A force that is becoming harder to contain each day. It was like trying to stop a tsunami with your bare hands. To the extent that it had the potential to overwhelm and even wreck the entire organization.

This is when I realized that data can easily create a Catch-22 situation—a difficult paradox with no apparent way out. Things can easily get out of hand if one does not know how to handle data the right way, and one can find oneself in situations that are often

impossible to salvage. Too dramatic? Well, let me take you through my journey over the years and I will let you decide for yourself.

Recognizing the power of data and analytics at McKinsey

My data journey started as a consultant at McKinsey in 1996. Those were the days when email accounts were not commonplace or easily accessible and mobile phones were a luxury. Yes, I know it is hard for the current generation, which probably grew up streaming cartoons on YouTube, etc., on their parents' mobile phones, to believe life existed without phones. My son, who is a product of the current generation, found it astounding and asked me once, 'How did you contact someone in case of an emergency?' Well, we did, and we did okay. But I am sure most millennials and all baby boomers can relate. Anyway, I digress.

So, as a consultant at McKinsey, my main job was to crunch data; 1996 was a data-scarce world, and McKinsey was a leader in strategy consulting owing to its capability to dig out data from nowhere, make sense of it and package it for clients. And that's what people like me, from the best business schools of the country, were hired to do.

For example, if we had to do a tractor market sizing in Brazil, believe it or not, it used to be a three-month-long project, because such kind of data wasn't readily available anywhere in the pre-internet era. And this was literally a brain-bending exercise. We would spend days collecting information from various global locations, often travelling across multiple countries, then collate it and size the market based on multiple approaches. To build confidence in the analysis, we used methods like triangulation—using three different approaches to get to some estimate and then narrow it down to a number that appeared to be the most logical. The catch here was that no matter how logical it was, it was still nothing more than an educated guess. And some of the best minds were employed to do just that. Such was the reality of the data-scarce world.

At McKinsey, my specialization was operations diagnostics and benchmarking, travelling all over the world, comparing a company's internal data with external best practices and extracting insights to highlight improvement opportunities for the company. So, I was dealing with data and analysis day in and day out. After four years of doing that, I left the company to build my own start-up for the next two years. But as fate would have it, things didn't work out as planned and I got an offer to join back at McKinsey to head the McKinsey Knowledge Centre (McKC) that was set up in 1998. At that point, it was quite fledgling, with just a few people trying to make something of it. My job was to lead this effort and to transform this back-office kind of set-up that provided overnight research support to a few teams in North America into a true knowledge centre. But the firm didn't have any concrete plan to achieve that.

When I came onboard, I realized that just in the span of two years that I was away from McKinsey, from 2000 to 2002, an *information revolution* had happened. I was literally blown away by how much had changed in the way data was now being generated and consumed. And the key force behind it was the explosive growth of data because of 'the Internet'. And within a short span of two to three years when I wasn't looking, the world of information had completely transformed. The information that we toiled for three to four months to gather, something that required highly skilled, highly paid professionals to be employed, was now available at the click of a button and could be done by anyone who had internet access and basic Google search skills.

This was a hard-hitting moment for me. A moment of revelation that McKC, which so far was considered the poor cousin of consulting, suddenly had the keys to the treasury. It was now the fuel that fired the consulting engine and had the ability to either empower or dis-intermediate the consulting teams. We had very quickly moved from an information-scarce world to an information-abundant world. I saw the opportunity that came with data abundance. Data became a source of knowledge that could be translated into insights very quickly and effectively. Voila, I had

found the secret ingredient to pivot McKC into a one-of-a-kind, best-in-class knowledge centre. How?

Let's consider the tractor market sizing scenario again. Now, data was available on not just the market size, but also the number of players, market share of every player, etc. All the information was just a mouse-click away. In addition to that, since this data kept growing, we now had the opportunity to capture it both longitudinally—over several years, and cross-sectionally—across multiple players. So, the role of analysts evolved from searching data to bringing it all together and analysing it to generate deeper insights. The Knowledge Centre had become more than just an information dissemination centre. It was now a Research and Information Centre. We started realizing the value from data because now we had enough information that we could do research on top of it. There was no stopping us now. As the research capabilities grew bigger and more sophisticated, the Research and Information Centre was renamed the 'Knowledge Network'. And as data continued to grow, we started to bring it all together from various industry verticals and functional areas in a systematic manner, which became the foundation for what we call 'Proprietary Knowledge' and the creation of proprietary databases at McKinsey. As a result, the expertise that was earlier in the minds of people could now be institutionalized. For example, now we were able to develop diagnostics that would systematically bring the benchmarking data for each industry vertical together, which is the starting point for many projects that McKinsey takes up today. We were able to build deep expertise and research capabilities that truly transformed the way one of the world's best consulting firms engages with its clients today. This is how we graduated from data to analysis to insights and finally to truly generating value from data.

And that brings me to another interesting story about setting up analytics capabilities at McKC. In the data-scarce world, problem-solving was predominantly done using deductive logic—the hypothesis route. But now suddenly we had access to bytes and bytes of customer data that could be used to do precise segmentation. And this data was growing by the minute. Suddenly

we were dealing with not just ten but 10,000 data points. Analysing data on such a large scale required technologies more advanced than Excel. And lo and behold! We had technologies like statistical analysis system (SAS) and statistical package for social science (SPSS) then, which could deal with up to 2 billion rows of data at a time.³ This was another eureka moment for me as I recognized a whole new world of opportunity that these technologies could open for us. We no longer had to rely on intelligent guesstimates. We could analyse large sets of data faster, with less manpower and greater accuracy and perform varied types of analysis that we had never imagined possible. I also realized that as data kept growing, it would become impossible to generate comprehensive insights without upscaling to newer and more sophisticated analytical tools, and we could face the risk of being left behind as a consulting firm.

Often true with any mega trend, even the best and the most experienced minds may not be able to foresee the potential of an idea before it becomes a phenomenon. Also, moving away from the tried and tested formula that has worked for years, to try something new and radical isn't that easy for all; this was also the case in a great firm like McKinsey at that time. But I was convinced that setting up a specialized analytics capability at scale was, in fact, an important pivot point for McKinsey. So, I pitched it several times, before multiple stakeholders at McKinsey, including going to the Shareholders Council, the highest decision-making forum within McKinsey, with a detailed study and numerous proof points. Yet, it was difficult for the top brass at McKinsey to fathom that this new breed of analysts sitting in India could achieve, or maybe even be better at, something that the best minds from the likes of Harvard could not do. Then again, I wasn't ready to give up yet. Even at the McKinsey Practice Olympics, the firm's initiative in 2005–06, where colleagues would submit proposals for new knowledge and innovation initiatives, I presented my perspective on how analytics was going to change the face of consulting in the future. Alas, there too I was met with a lukewarm response. Undeterred in my conviction, I continued to pursue the idea. In the absence of top-

down support, we took the route of bottom-up experimentation, and this opportunity certainly existed in a decentralized global firm like McKinsey. It took a while, but we did end up building a series of specialized analytics capabilities, most of them completely new for McKinsey, and helping it create new areas of engagement with its clients. Analytics and other data-driven capabilities that I mentioned earlier helped not just create the most innovative knowledge centre of its time, but fundamentally reshaped how a great consulting firm like McKinsey serves its clients. Today, digital and analytics, the combination of proprietary knowledge and the platforms, make up for a pretty big chunk of McKinsey's global revenues.

That is the story of how data and analytics, in just a short span of twenty to twenty-five years, changed the consulting landscape, impacting it in such profound ways. A firm like McKinsey never really planned for it initially, and many organizations like them, owing to the sheer force of the changes that data was bringing about, had to completely transform their business models to adapt, survive or stay ahead of the curve.

Dealing with the dark side of data at Flipkart

Now that I have talked about the good side of data, let me also unveil the dark side of data. This story began when I joined Flipkart in early 2016, which is one of India's dominant e-commerce giants. But back then too, it was considered a poster child of the Indian start-up explosion. And the culture at Flipkart was markedly different from what I was exposed to. While McKinsey had more of a hypothesis-driven approach, Flipkart was absolutely obsessed with data. Everything was data-driven, with every employee exposed to and dealing with thousands of data points every day. Over 100 people joined our daily stand-up meetings, and everyone talked in numbers. And I, an IITian, an IIM grad, an ex-McKinsey consultant, a chief experience officer (CXO), who considered myself pretty good at data analytics, was struggling with the endless sea of numbers.

Our biggest project at that point was the Flipkart Data Platform (FDP), to bring all the Flipkart data on to one platform, creating a

single source of truth and building a single view of the customers. Flipkart was obsessed with becoming truly data-driven by bringing all the information to one place to be viewed equally by everyone working there. We believed that this was the answer to generating unique and innovative insights. Now, between 2014 and 2016, Flipkart was on an acquisition spree. It acquired Myntra and Jabong, two e-commerce companies, to add to its fashion and lifestyle segment and acquired a stake in MapmyIndia to enhance its delivery operations. The year 2016 was also momentous for Flipkart as it crossed a significant milestone of acquiring 100 million customers.⁴ With such a significant customer base and expansion in operations, imagine the amount of data that was being generated at that time. As a result, the situation quickly spun out of control. By the end of 2016, Flipkart was going through the biggest crisis since its inception. As a result, a major internal shuffle ensued at the top management level, Amazon was slowly invading the e-commerce space, the company was losing more than a billion dollars every year and was on the brink of becoming cash-strapped. In essence, the \$7-billion company was literally coping with an existential crisis at this point. But the biggest problem the management team was often debating was not that we were losing money, or that Amazon had beaten us, or that we could not find investors. The biggest problem for us was that the data platform had become a complete bottleneck, which significantly impaired decision-making and execution in the company.

A single source of truth and a single view of their customer was expected to act like a turbocharger, boosting us to dash ahead in the e-commerce race. But contrary to that belief, it ended up stalling the company's decision-making and operations, bringing everything to an abrupt halt. The single source of truth now turned into the biggest roadblock for the company.

Here is the harsh reality. The one key error in judgement that most organizations and their data and technology experts are making is underestimating the pace at which data is growing and will continue to grow in future. As I explain later in Chapter 2, data

is growing in geometric progression whereas technologies that are being built to deal with this data are growing in arithmetic progression. So, the gap will continue to widen, with technology always failing to catch up.

Therefore, learning it the hard way, I have now ceased to believe in the myth of a single source of truth or a single view of the customer. It is a very difficult feat to achieve.

In Flipkart, we realized it the hard way too. The best engineers of the world—Flipkart in those days hired the crème de la crème—got down to solving it. But alas, ‘all the king’s horses and all the king’s men couldn’t put Humpty together again’. We realized that we must let go of this impossible task and tackle the data problem in a completely different way. Luckily, we had some practical minds who could step away from the endless technology puzzle that we were trapped in. And that’s when we realized that we could not solve the data problem only as an engineering problem. It required a radically different approach. Since the data was too large and varied, quality was always an issue. And the Flipkart business was dependent on taking real-time actions based on real-time data. The challenge was, how do we get the right data and accurate data each time? And this was the moment when another myth of Big Data—that the law of large numbers will take care of quality—was also dispelled for me. It did not work in real life. It was an impossible task to manage the quality of the truckloads of data that was pouring in every day.

The answer was to start small. We narrowed down the problem to make it more manageable. We decided to first tackle just four critical decisions—we called them use cases that we wanted to drive. To drive these decisions, we identified the number of data pipelines required for just those, which was around 100 versus the 7000, which we were dealing with earlier. These pipelines were then cleaned and curated for those use cases, and since this was a manageable task, we could make it work. And then we continued to replicate the same mode across a greater number of use cases as well, and slowly and gradually we dug ourselves out of a very big ‘Data’ hole.

And the data journey continues

My data journey continued as I joined as the CEO of Incedo Inc. in late 2017. And believe it or not, I was flabbergasted to see that one of our biggest clients, a global financial institution, was also falling into the same trap of trying to create a single source of truth, by consolidating all their data from a risk and compliance perspective. Hundreds of millions of dollars had already been spent on this project over the years. But eventually, the project was scrapped since they were running after the unachievable. I stepped on the scene a bit too late to stop the disaster from happening. But here was another proof that such big bang approaches rarely work and that is the reason why Big Data initiatives that cost a lot of money still often fail to deliver the expected impact.

In contrast, another client of Incedo, a telecom player, was trying to bring all their customer data together to create 'a data lake of all data lakes'. And I could clearly see them rapidly cruising towards a dangerous whirlpool of never-ending spend on data infrastructure build-out. But here we could help salvage the situation. We did a lot of groundwork and built a case to prove that it would be better to take a more focused, business use-case-centric approach. Based on this, the mega infrastructure build-out that was being envisaged was moderated.

Key themes covered in this book

This book on data is my effort to bring forth the learnings from such experimentations and suggest actionable strategies to harness the full potential of data, especially as we march into the AI age. I do not claim to have found the holy grail, because data is such a vast and complex topic that is becoming even more complicated and therefore has not been fully understood. But over the years, I have worked tirelessly to understand how data works and believe I have succeeded in uncovering a few key mantras to unleash the power of data. Through this journey of exploration and revelation, I intend to provide some critical insights and a structured approach to

organizations, entrepreneurs and young professionals alike. And this book is not just for those who deal with data day in and day out. There is something in it for everyone. From the individual who is interacting with data every day through the internet to multi-million- or multi-billion-dollar companies that are in the midst of their digital transformation agenda to political leaders looking to solve complex national and global problems. Data is a valuable asset for everyone, and so dealing with it is a critical skill that each one of us should master to win in the data-first world.

The book is divided into three sections. **Section I** is dedicated to **Understanding the Data-First World**. The fuel powering the digital age and the upcoming AI age is data—like electricity powered the second Industrial Revolution,⁵ in this fourth Industrial Revolution (the digital age), data is a source of significant opportunity and advantage if utilized the right way, as proven by digital natives like Google and Amazon. This section lays out the opportunities that the explosive growth of data has brought about for organizations and individuals and delves into the challenges that emerge as organizations are now forced to deal with exponential amounts of data, while the technologies are playing catch-up.

This section has five chapters. I begin by highlighting the **Data Explosion** in the digital age (Chapter 1). With the advent of the internet, data witnessed exponential growth, leading to the phenomenon of Big Data. This growth evolved across three core dimensions, volume, variety and velocity, as data expanded in terms of size, type and speed of generation. Then I put a spotlight on **Data, the Fuel for the Digital Age**, that powers every aspect of life and business, transforming experiences and pushing industry boundaries (Chapter 2). As organizations gain access to vast customer data, they are able to unlock unprecedented opportunities for value creation and reimagine business models. Next, I highlight the possibility of **Value Reimagined** through the data-insights- actions-impact (DIAI) framework, enabling organizations to realize transformational value from data (Chapter 3). By generating insights

and taking effective action, organizations can deliver tangible positive outcomes for customers and themselves.

This raises an important question—despite the ubiquitous nature of data and the unprecedented opportunities made available, why are most organizations struggling to maximize value from data? My study reveals the contradictory nature of challenges that most organizations face today—**The Data Paradox** (Chapter 4). On the one hand, they are unable to deal with the overwhelming 3V (volume, variety, velocity) explosion of data, while on the other, they struggle to get relevant insights to enable data-driven decision-making. Despite continuous attempts to deploy innovative technologies, the problems persist. On closer examination, I discovered **The Root Cause** to be a disproportionate focus on solving the data problem with technology and infrastructure, which in my experience, is not enough (Chapter 5). Most times, the root cause of the data paradox is more logical (narrowing the problem to be solved) than physical (technology infrastructure), and can have a wider impact on the organization, if it remains unresolved.

Thus, to navigate through the choppy waters of the data revolution age, organizations must understand the root cause of the problem to narrow down the issues and attempt to solve it through a structured and comprehensive approach.

In **Section II, Maximizing Value in the Data-First World**, I propose an innovative and practical way to make sense of all the chaos brought about by the explosive growth of data to maximize its value realization. I propose a thirteen-component solution framework which I call the **Unified Solution Framework** that can very well be the thirteen mantras for organizations to succeed in the data-first world and the AI age.

These thirteen components are divided into five layers. Layer one is the **Business Objectives**, which is the starting point for any organization. Here, I emphasize that it is critical to clearly **Define the Business Problems** (Chapter 7). Because in my experience, breaking the business problems down into logical components, prioritizing them and mapping data requirements accordingly, in

effect narrowing the data problem to be solved, is the only way to make the data process more focused and manageable.

The second layer is recognizing and understanding the rich, diverse **Data Ecosystem** made available to organizations today because of the 3V explosion of data. Here I have highlighted three key types of data—multi-source data, real-time data and proprietary data (Chapters 8–10)—which I believe drive disproportionate value generation for organizations. **Multi-Source Data** is critical to generate comprehensive and more meaningful insights. **Real-Time Data** is essential to drive action with speed and **Proprietary Data** enables organizations to build sustainable competitive advantage.

As I mentioned before, despite the abundance of data available today, organizations struggle to leverage it effectively due to increased complexity and scale. Keeping up with the volume, variety and velocity (3Vs) of data and translating it into meaningful insights or actions requires a well-thought-out data architecture. So, the third layer, **Technology Infrastructure**, emphasizes the right way to bring the 3Vs of data together, in a form where it can be used in combination to solve multiple business problems. Here I emphasize the criticality of building a customized data stack that enables scalability and facilitates end-to-end integration of the various layers of the data management value chain, which isn't possible without building a **Modern Data Stack** (Chapter 11).

Once the data ecosystem and the technology infrastructure are established, organizations must build **Core Processes**, which is layer four, to achieve the desired returns from data initiatives. Setting up and institutionalizing these core processes is important for the successful implementation of any data initiative. Here I highlight the importance of **Agility** to successfully adapt to the evolving nature of problems in a VUCA (volatility, uncertainty, complexity and ambiguity) world (Chapter 14). I recommend a 'two-speed approach' to start by delivering on short-term business value through Speed One initiatives, while building long-term capabilities, Speed Two, by bringing together Speed One initiatives. Furthermore, I discuss the importance of **Data Democratization** (Chapter 15). Making data available to all employees appropriately with seamless, any-time

access while empowering them with the right level of knowledge and know-how to use it effectively, is critical to enabling better, more informed decision-making and driving actions faster across an organization.

And while it is paramount to make data accessible and available to all, all the efforts towards building a data-driven organization could easily be put in jeopardy if the right **Data Security** measures are not taken (Chapter 16). And with the evolving complexity of the data threats, and the growing vulnerabilities brought about by ever-expanding data and technology ecosystems, I recommend implementing a zero-trust architecture, keeping the context in mind.

And finally, **Organization and Culture**, which make up the fifth layer of the Unified Solution Framework, address the mindset and behaviour of people in a manner that fundamentally transforms the organization's DNA, thus impacting its overall ability to drive transformational value from data. Here I talk about the need to rethink organizational structures, decision-making processes, and the most effective organization design to execute data initiatives for building effective **Organizational Alignment** (Chapter 17). I also emphasize the importance of adopting a three-pronged approach to move away from gut-based decision-making, where decisions are taken based on the highest-paid person's opinion (HiPPO) and build a robust **Data Culture** (Chapter 18). And finally, as the Big Data world becomes more complex, technologies evolve rapidly and AI goes mainstream, I see the role of specialized **Data Talent** evolving significantly as well (Chapter 19). Data talent needs to move beyond the core data skills to include aspects like deep problem-solving, critical thinking, creativity, storytelling and deeper domain knowledge, as the focus shifts to achieving business outcomes and driving impact.

The five layers that must come together to create a data-driven organization require something to hold them all together. The glue that unifies the efforts at each layer of the Unified Solution Framework are two critical elements that I call the **Integrators**. The first one, **Data Quality**, to me, is the biggest concern of the Big Data world, as data continues to become more complex with

huge and diverse data sets being continually generated and new use case needs emerging (Chapter 12). I propose a context-first approach to tackle quality issues. The second integrator to me has the potential to become the most effective solution to maximizing value from data. These digital business solutions, called **Data Products**, built as a vertical slice integrating various elements across the data stack to deliver a business outcome, can help organizations deliver impact at speed and in a repeatable manner (Chapter 13).

In **Section III, Data for Individuals and Beyond**, leading in with the quote, 'As is the microcosm so is the macrocosm, as is the macrocosm so is the microcosm', I change gears and zoom in to the individual level and zoom out to the macro level of society and nations. I explore the universal nature of the data paradox, highlighted in Chapter 4, which plagues organizations today. I emphasize how the data paradox that is true at the enterprise level also exists both at the individual level and the society/global level (Chapter 20). I start this section by elaborating on how Big Data has brought about **The World of Hyper-Personalization**, creating a 'segment of one', where every individual is unique and therefore the products and services are increasingly being customized at the individual level and not at the mass level as was happening traditionally (Chapter 21). But for this to happen effectively, individuals need to engage more with the digital world, allowing organizations to better understand their customers and generate greater value through personalization.

Then I talk about how, despite so much data being available to individuals, most of us are still relying on the gut-based approach to make decisions in real life, whether big or small. I find the DIAI framework, which I talk about in Chapter 3, useful at the individual level as well to effectively use **Data for Better Decision-Making** (Chapter 22). Furthermore, while access to data is valuable, it can lead to information overload for an individual and make it challenging to extract meaningful insights for decision-making. The key is to balance **Information and Wisdom**, by building wisdom,

which is proprietary to an individual, to cut through the noise by identifying patterns that matter most (Chapter 23).

Moving on to the paradox of **Digital Engagement vs Mental Health** (Chapter 25). Information overload leads to noise that surpasses an individual's capacity, creating mental havoc. The solution lies in finding ways to create a stronger connection with oneself at a deeper level and to get to the right balance, which to me is spirituality. I further move on to address another paradox of the data world, **Data Sharing vs Data Privacy**, which is becoming a greater concern as we continuously engage with the digital world, generating and sharing data about ourselves every minute of every day (Chapter 24). While it is difficult to keep track of how and which data is being accessed for which purpose, we must be vigilant in protecting the few critical data points most vulnerable to misuse.

Now moving beyond individuals, I believe in the importance of **Data Collaboration for a Better World** (Chapter 26). Global data collaboration can address complex problems like climate change and healthcare that no one country can solve alone and help create a more equitable and sustainable world. And finally, I have no doubt that a new world order will emerge as the AI age unfolds and nations that use **Data as the Source of National Competitive Advantage** will emerge as the new superpowers (Chapter 27). So, data is a topic not just for enterprises and individuals, but for national leaders and policymakers as well.

The pages that follow are the result of my many deep encounters with data over time. The more I have experienced and researched the way data works, the more convinced I got of its boundless potential to generate extraordinary value and that now is just the beginning. Through this book I lead you into this remarkable world of data to open your eyes to its full potential, truly making it the pivot that can help organizations win in the AI age. But as data continues to explode in all directions, the data paradox is overwhelming most organizations, leading to frustrations and challenges in value creation. So, through the Unified Solution Framework that I present, my aim is to provide every individual,

professional and organization with a playbook to succeed in this data-first world.

I have endeavoured to delve into every component of the solution framework, attempting to demystify each concept in detail and highlighting its importance in the data-first world and its growing criticality as the AI age unfolds. As I have mentioned earlier, data and AI have a recursive relationship, so AI would play a critical role in solving each of these components. At the same time, solving these components would be critical to unlocking the full potential of AI. In simple terms, the better organizations get at dealing with data, the more value they will be able to generate from their AI efforts and AI will play a significant role in solving some of the key issues organizations face with data. I have attempted to provide a structured and practical approach on how to assimilate each of these components into the DNA of an organization to maximize value from data in the data-first world, an approach that is industry- or function-agnostic. I have also called your attention to the fact that this data paradox exists at both the individual and macro levels and that various components of the solution framework can act as a guide to resolve the paradox at those levels as well.

It is truly exhilarating to find ourselves at the forefront of not just one, but two remarkable phenomena that are unfolding almost simultaneously. The first is the emergence of a data-first world, where data has become a central driving force, shaping industries and fuelling innovation. The second is the dawn of the AI age, propelled by the advent of generative AI, which has created the possibility to leverage the data of the world for the first time. The convergence of these two holds immense promise and the opportunities are boundless.

And with this book, I intend to equip you with both the principles and practical frameworks to tackle the challenges of the data-first world and get ready for the AI age.

With that thought, let's embark on this fascinating data journey together!

Through this book, I will attempt to answer the following key questions:

1. Why and how can AI powered by data create transformational value for enterprises and individuals?
2. The world is full of paradoxes and so is the world of data. How can individuals and enterprises effectively deal with the paradoxes to unlock the transformational value of data and AI?
3. Are there some principles that we can learn from life and apply to data and vice versa, as we navigate the new world of data-enriched lives?



SECTION I

UNDERSTANDING THE DATA-FIRST WORLD

'Water, water everywhere, nor any drop to drink.'

—*Samuel Taylor Coleridge*

Data Explosion

An Unprecedented Phenomenon

'Sometimes you will never know the value of a moment until it becomes a memory.'

—Theodor Seuss Geisel aka Dr Suess,
American writer and cartoonist

Data has been an integral part of human evolution and rapidly evolving technology has enabled unprecedented growth in data being generated and captured. Data has been growing and evolving at a trajectory unfathomable to any individual or organization. And just like that, in just a few years, we find ourselves entering a new era, an era of data—the data-first world. As we are presented with countless opportunities to fully exploit the potential of data, we must ride the wave before it's too late. We must find ways to unlock transformational value from data now and in the years to come.

The dawn of the data-first world

As we are living in what is aptly called the digital era, the way we live, work and play have all transformed dramatically. And all this has happened in a noticeably short span of twenty-five years—just a quarter of a century. Digital has affected our lives in ways unfathomable and will continue to do so in the years ahead. This digital age is unique I believe, it is so revolutionary, impactful and all pervasive, that it is impossible to miss it. We all are a part of it, experiencing it, and benefiting from it directly.

There are *six key characteristics of the digital age* that I frequently talk about. Let me quickly recap these for you:

1. **New-age customers:** They are trending younger, have higher expectations, are more demanding as they have more choices available to them. They are tech savvy and highly active on social media which is also a key influencer in their decision-making process.
2. **Technology at the core:** It is becoming the core of every business. Technology started as a support function, it became an enabler along the way, and today it is front and centre—forcing every organization to reinvent and build a technology DNA.
3. **Velocity of change:** The velocity of change is accelerating on all fronts. Technologies like cloud and mobile have become commonplace, replaced by newer technologies like AI, IoT and blockchain driving conversations today.
4. **Iterative approach:** Constantly evolving customer needs, explosion of data and rapidly maturing digital technologies are moving us from linear to iterative approaches or continuous experimentation—taking quick actions, assessing results, learning and tweaking and re-executing.
5. **Interdisciplinary issues:** Digital problems are cross-functional, requiring seamless collaboration across many functions—finance, HR, marketing, sales, operations, engineering working collaboratively to rapidly develop solutions.
6. **Explosion of data:** The volume, velocity and variety of data generated today are unprecedented and will continue to grow at an exponential rate. This phenomenal growth is impacting individual lives and businesses in dramatic ways, presenting every business opportunity to use it in new and innovative ways.

Of these six key shifts, *the most transformational one, according to me, is the explosion of data—the data revolution.* The digital transformation spanning across just a quarter of a century that led to an explosive growth of data has gotten us to the **data-first world**. With the rapid growth and adoption of digital technologies, the amount of data created has grown exponentially. Most of the world's digital data has been created in the last three to four years and is expected to *double by 2025*. Let me give you a minute for that fact to sink in—*most of the world's digital data is merely three to four years old and it will double in size in the next three years.*¹ It is a real-world example of the classic 'hockey stick' growth rate curve that is cherished by venture capitalists and executives the world over—a long period of relatively stable, minimal growth followed by a sudden and consistent steep upward trajectory of rapid growth.

Today everything in the digital age, from humans to machines, is a **data factory**—every time a human uses digital or connected devices or gets on to the internet, he/she generates data points and every machine hooked on to the IT system generates data. And with such phenomenal growth, the potential and possibilities with data are endless. It has become an amazing and powerful force that is transforming our lives and businesses continuously.

The good news: it is happening now, and it is just the beginning!

I believe we are at the very early stages of the data revolution. We have just set foot into the data-first world. We haven't even scratched the surface in terms of realizing the full potential of data. And the leaders, visionaries, companies and societies who will create the future have still not entered the mainstage—your opportunity to lead, change and make the world better is still readily available for your role in this data-first world.

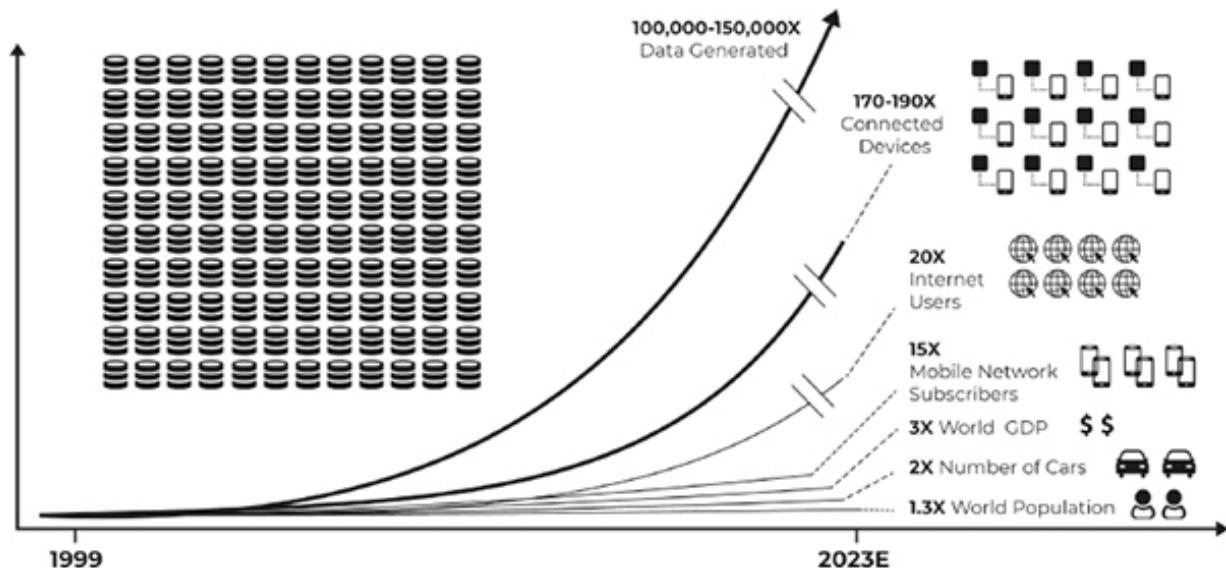
How do I know this? In the 5000 years of recorded human history, most of the data has been created over the past 3–4 years. And within this short time span, it has witnessed unparalleled growth.

But it is still a very new phenomenon for anyone to fully comprehend. What we know for sure is that it is already creating enormous value, as is made evident by the fact that four out of the five most valuable companies in the world (Apple, Amazon, Google, Microsoft and Walmart) are essentially data companies² that have designed their business model around data to create value in unimaginable places! However, for most organizations, data is still a big challenge, an enigma that hasn't been figured out. Therefore, this is just the tip of the iceberg. We are still in the early stages and the possibilities are endless. And while data has resulted in some radical changes around us, it still has a potential to make wide-reaching impact and bring some universal changes in the world around us.

Furthermore, every revolution leads to changes and shifts in power at multiple levels, from regions, nations, companies, groups, to individuals. So be prepared for another change in the global power dynamics driven by the data revolution. From the global power the Great Britain possessed during the age of steam power, to the United States' emergence at the dawn of the nuclear age, history has repeatedly demonstrated that as big revolutions lead to big changes in the world order. The data-first world will be no different, where the early bird will catch the worm. The leaders of the new world order will be determined based on the ability to foresee the potential of data and leverage it effectively to reimagine value, before everyone else catches up.

This is why I believe we have already entered the data-first world, a revolution where data not just facilitates but fuels the transformation across all fronts. Success now depends on the differentiation created by maximizing the value realized from data, especially as the AI age dawns upon us. This epochal shift hinges on one powerful component: **Data**.

Data has exploded at an unprecedented rate over the past two decades and is the most defining trend of Digital Age



A phenomenal growth story

Yes, I keep saying that data has seen unprecedented growth over the past few years. So what? One can counter that internet users have grown significantly too and so have mobile users. What makes the growth of data specifically 'phenomenal'? So, let's put it in perspective, shall we?

Let's compare the growth trajectory of some important consumer and technology metrics that happened over the past twenty-four years. Why 1999? Because that's when the first comprehensive study was done to quantify the total amount of digital information (data) available in the world.³

Between 1999 and 2023, the total human population on our planet grew **1.3 times**, from 6.07 billion to 8.05 billion.⁴ The world GDP **tripled**, from 32.9 trillion US dollars to 105.5 trillion dollars,⁵ brought about by globalization, technological advancements, rapid growth in emerging markets, infrastructure investments and many other factors.

The virtual world saw much faster growth during the same period. The period from 1999 to 2023 marked a significant and rapid

expansion in the number of mobile subscribers globally. Fuelled by technological advancements, affordability and improved network infrastructure, mobile subscriptions grew more than **fifteen times**, from 0.5 billion subscribers to 8.3 billion in 2023.⁶ During the same period, with the advent of the internet, the world saw a profound shift in how people around the world connected, communicated and accessed information, resulting in a staggering **twenty times**

growth in internet users from 248 million to 5.3 billion.⁷ This period also saw an IoT revolution and technology miniaturization, which led to a seamless integration of connected devices in our lives, through smart devices, smart homes, cities and industries. This network has grown almost **170–190 times** bigger, from mere 90 million devices to a network of more than 15–17 billion connected devices.⁸

While these numbers are huge, there is one phenomenon, which lies at the heart of all these advancements—data. And as various technologies emerged and their adoption picked up pace, the data that powers it all grew quietly and unassumingly. While we were applauding the remarkable growth of internet users and mobile phone subscribers, data surreptitiously grew by, not thirty, not 100, not even 1000, but a whopping **100,000–150,000 times*** give or take!⁹ We will be hard pressed to find similar examples of such explosive growth in such a short period of time across the vast expanse of human history! Now, isn't that phenomenal?

Let us look at the history of this phenomenon.

Evolution of data

Webster's Dictionary, or more appropriately since we are in the digital age, www.merriam-webster.com defines data as 'factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.' True. However, I would add one more element in the context of this book. It is critical that data and the corresponding analysis also give you a basis for 'action'.

The earliest evidence of data goes back to around 18,000 BCE when palaeolithic tribes would mark sticks and bones with notches to estimate how long their food supplies would last.¹⁰ Another notable milestone in data history is the use of data by the Romans. The Library of Alexandria, dating back to 300 BCE, is probably the earliest attempt by ancient Egyptians to gather all the data related to their empire. The Romans, seeking accurate census data for taxation, introduced 'census takers' who recorded and centralized information from diverse regions.¹¹ The Romans also pioneered statistical analysis for war strategies, to predict likely insurgent borders and deploy their vast army of 5,00,000 soldiers in the most efficient way, ¹² across a sprawling empire spread across 50,00,000 square kilometres. ¹³

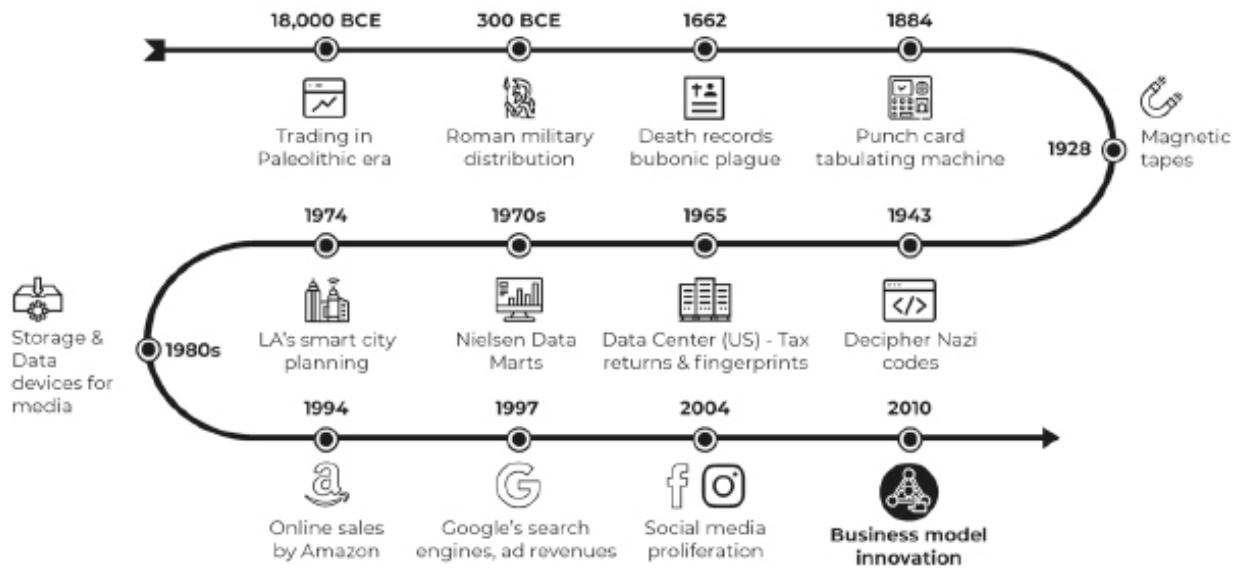
Fast forward to the seventeenth century, in 1662, John Graunt introduced **statistical data analysis** when examining the tragedy of the bubonic plague.¹⁴ The London haberdasher (what we would consider a men's clothing retailer today) published the first collection of public health records. During the bubonic plague in England, he compiled and analysed church records of recorded death rates and their variations.

In 1884, Herman Hollerith invented the punch card tabulating machine, marking the earliest attempt at **data processing**. In the 1880s, the US Census Bureau used this machine to measure and record the population growth, saving the government US\$5 million. Hollerith later founded the Tabulating Machine Company in 1896, known as IBM today.¹⁵

The genesis of modern data storage, known as **data centres**, dates back to the 1940s with the creation of the world's first programmable computer, ENIAC (the Electronic Numerical Integrator and Computer).¹⁶ Initially designed for complex artillery calculations during World War II, ENIAC was the first programmable general-purpose electronic digital computer that could follow different instructions or change their order based on certain data values. But

due to its limited twenty-word memory, a parallel data storage system had to be developed to bolster it—a peak into today's 'hyperscale data centres' established by major global tech firms.

Data has been an integral part of human evolution and its impact has increased exponentially over the years



By the twentieth century, the pace of change had hastened as technology spurred major developments. As technology evolved, the complexity around handling data grew significantly. For example, organizations were not just dealing with transactional data generated internally, but also needed to integrate external databases and applications that were expanding rapidly. As a result, the data was poorly integrated and inconsistent. There was a dire need to find a more scalable solution to integrate such diverse data sources in an effective way to facilitate decision-making. **Data warehousing**—a way to bring data from many different sources together to a single location and translate it into a format that is easy to process and store—emerged as a solution. ACNielsen provided clients with something called a 'data mart' in the early 1970s where the end users—the retailers—had direct access and control over their data. This condensed and more focused version, the data marts, laid the foundation for data warehousing solutions,

popularized by IBM in the 1980s, through its business data warehousing offering.^{[17](#)}

The civic authorities weren't lagging either. Cities had already started gathering and using civic data in the 1960s. Los Angeles's Community Analysis Bureau was already using computer databases to run cluster analysis—the process to find similar groups of objects. By combining this with computer databases and infrared aerial photography, they created reports on neighbourhood demographics and quality of housing to direct resources to tackle diseases and poverty.^{[18](#)}

In 1971, IBM introduced floppy disks that were widely used from the mid-1970s to late-1990s. The initial read-only versions had a storage capacity of a mere 79.9 KB. In 1982, Sony and Philips' CD player recorded 60 minutes of audio/video. By 1999, the SD card, by Panasonic, Toshiba and SanDisk, could hold 8 MB. And over the years, physical storage space shrank while capacity grew. Today, a nano memory card can store 256 GB of data. We're moving towards limitless virtual data storage.

The next significant development was Amazon's foray into e-commerce, beginning as an online bookstore. By 1998, Amazon's expansion beyond books marked a turning point. Pioneering collection, storage and analysis of data on their business, customers and ecosystem partners, they introduced innovative features like one-click ordering, prime subscriptions, 360-degree customer profiles, etc. In 2006, Amazon web services (AWS) revolutionized computing with cloud infrastructure, enabling remote data processing.^{[19](#)} Currently, AWS dominates the cloud services industry with roughly one-third of the global market share.^{[20](#)}

In parallel, in 1996, the epic internet search engine, Google, was launched by Larry Page and Sergey Brin on Stanford University's network. Initially called BackRub, Google began as a research project.^{[21](#)} It was conceived as a system that would crawl the internet to determine which pages were linking to other pages. It became the foundation of the largest search engine in the world that brought

data from every part of the world to a common platform, making it discoverable by any user who typed in the keywords in the search bar. Today, it is often flippantly equated with the omniscient God. Google knows it all!

The internet also opened ‘windows’ (pun intended) for people to connect with each other, regardless of the distances that physically separated them, giving rise to another phenomenon, social media. Social media platforms like Orkut, not many from the new generation might know something like this existed, Facebook, Myspace, Twitter (now X), YouTube, were all founded around the same time between 2003 and 2006. And within the next few years, the very meaning of social interactions and networking changed forever. And it changed the very nature of how an individual interacts with data. Now, each individual isn’t just consuming data, they are also creating it. With millions of posts, likes, photos and videos that people upload every day, social media has become a significant contributor to unstructured data being generated today.

Post-2010 multiple business models emerged that leveraged data to solve a customer problem or deliver unique, never-before offerings to customers. For example, the B2B (business to business) model—business sells products or services to another business, B2C (business to consumer) model—businesses selling products of other businesses to customers and C2C (consumer to consumer) model—one consumer selling to another. Online financial services emerged to provide a comprehensive solution to all financial requirements of a customer virtually, by consolidating data across multiple platforms and providing customized financial solutions based on customer buying behaviour and transaction patterns. Other innovative business models like food delivery services, online travel booking platforms, cab hailing services and many other business models have emerged over the years that leverage data through emerging technologies to deliver value to consumers.

The growth and adoption of digital technologies and the enhancements in data storage solutions spurred the expansion of data in all directions and in all forms, growing bigger and bigger every day.

When data got 'BIG'

The advent of digital changed the definition of data as well. Now data is also digital. Digital data can be defined as 'information that has been translated into a form that is efficient for movement or processing'. In simpler terms, data is information converted into binary form.

Traditionally, what we considered as 'data' was very limited. With a limited number of technologies out there, the possibility of capturing data was also limited. According to the traditional view, the characteristics of data were as follows:

- Data was restricted to discrete facts, statistics or information that was often numeric in nature. It was therefore easier to manage, store and use.
- The majority of data that was captured was in a structured format, especially since it was all numeric, and so it was easy to process using traditional data-processing software.
- The number of sources that generated data were also limited.
- There were standard sources to record and report data that had limited capability to generate insights owing to all kinds of data constraints.

But today the picture has completely changed owing to the exponential growth in the sources that generate data and a dramatic growth in the technologies available to capture, store and utilize it.

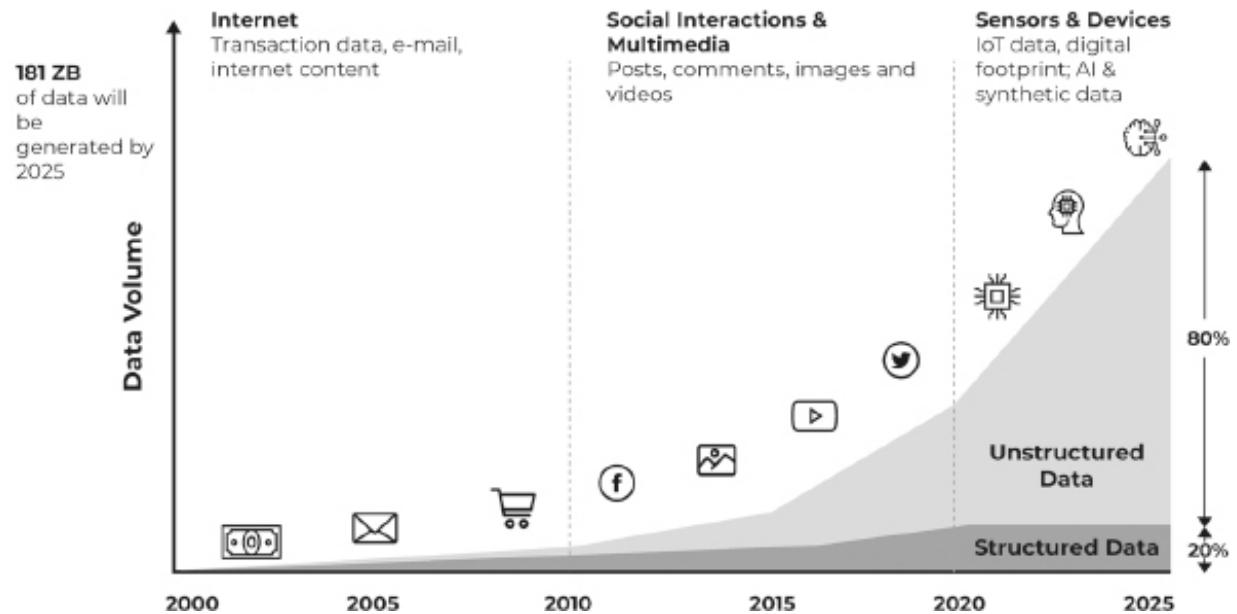
Enter ... **Big Data!**

The term 'Big Data' emerged in the mid-1990s. Although it is disputed who first used the term, most people credit John R. Mashey of Silicon Graphics for making it popular in 1996. In its true form, Big Data is not something new or something that originated in the last two decades. Immensely large data sets and systems and solutions developed to manage large accumulations of data existed before the digital age. But the true impact of it wasn't recognized in the general

public until it touched our daily lives via the likes of Facebook, Amazon, Apple, Netflix, Google—the FAANGs.

Especially since the beginning of the twenty-first century, the size, speed and variety of data generated have changed beyond measures of human comprehension. There is a large and diverse set of information that is being generated today at an ever-increasing rate. In addition to the structured data—typically characterized as quantifiable data that is highly organized and decipherable by machine-learning algorithms—that is steadily growing in volume and variety, we are bombarded with an unfathomable amount of unstructured data—mostly qualitative data that is available in free-form, less quantifiable and difficult to be analysed by conventional data tools and methods. It is estimated that around 181 ZB of data is expected to be generated by 2025,²² equivalent to 3.6 billion Blu-ray discs (with a capacity of 50 GB each). More than 80 per cent of this data will be unstructured.²³ This shift has been driven by the evolution of the internet, which not only added to the growth of structured data, but significantly impacted the expansion of unstructured data. The rise of social media amplified this growth further, expanding unstructured data manifolds. Adding to that, the recent advancements in technology have rapidly intensified this shift, resulting in an overwhelmingly large volume of unstructured data being generated and captured, as compared to structured data.

Growth of Big Data is predominantly driven by unstructured data



Complex Big Data sources like the Internet and IoT devices are multiplying the rate of data creation, accelerated by connectivity and cloud-based compute and storage. As a result, the traditional data-processing applications struggle to make sense of such complex and unstructured data.

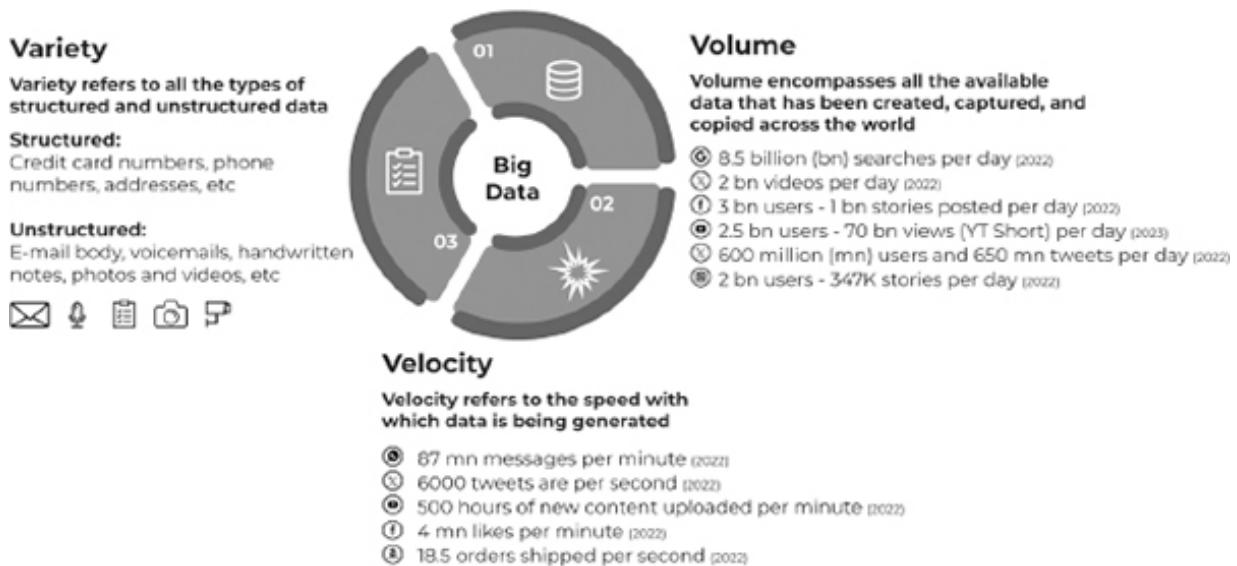
Such phenomenal growth of data cannot be attributed to just one or two factors; instead, multiple factors, like widespread adoption of digital technologies, increased internet connectivity, proliferation of smart devices, expansion of online services and more, have collectively contributed to this unprecedented surge. This growth has occurred on multiple dimensions, but for simplicity, we are looking at three core dimensions that broadly encompass the nature of the data explosion.

How data got so BIG—3Vs: The core dimensions

Given the varied and numerous forms of data, a seminal model was defined by Gartner analyst Doug Laney in 2001, which has stood the test of time. According to him, Big Data has witnessed a three-dimensional growth—the **3Vs**—in volume, variety and velocity.

But before we embark on the journey to understand the three dimensions, it's important to have a common understanding of how data is measured by most industries. *Data is measured using digital units like bits and bytes, reflecting how computer data has historically been stored.* The smallest unit is bit (b) representing a binary digit (0 or 1) and goes up to yottabyte (YB), the biggest unit, that helps us quantify and communicate the scale of data.

Big Data's explosion has happened across three core dimensions - The 3Vs



The byte sizes

- 1 byte = 0.001 kilobyte
- 1 kilobyte = 1024 bytes
- 1 megabyte = 1024 kilobytes
- 1 gigabyte = 1024 megabytes
- 1 terabyte = 1024 gigabytes
- 1 petabyte = 1024 terabytes
- 1 exabyte = 1024 petabytes
- 1 zettabyte = 1 trillion gigabytes
- 1 yottabyte = 1,204 zettabytes

It is interesting to see how the storage capacity of computing devices evolved over time. Back in the 1950s and 1960s, computers filled entire rooms but had limited storage capacity. The UNIVAC I, one of the earliest commercially produced computers, which was almost equivalent to a small shipping container in size, had a memory capacity of just 1000 words (roughly equivalent to 12 KB).^{[24](#)}

Then up until the 1970s, hard disks as big as a room were used for data storage, with a maximum capacity of 250 MB. In 1980, IBM introduced a more advanced hard disk, which could store up to 2.52 GB of data, almost four times of any IBM hard disk before. This one was significantly smaller, but was still as big as a refrigerator, and weighed 550 pounds (250 kg).^{[25](#)}

Fast forward to today, an average smartphone, small enough to rest in the palm of your hand and typically weighing around 150–200 grams (5.3–7.1 ounces), can hold anywhere between 64 GB and 256 GB of data—more than **100 times** the storage capacity of any computer of the 1980s!

Now, cloud storage, for which the physical servers are located at a data centre or server farm far away from the user location, has provided virtually limitless and scalable storage solutions! For example, based on estimates, total data stored in AWS is estimated to be around 1.4 ZB,^{[26](#)} which is equivalent to the storage capacity of **5.5 billion** smartphones (with 256 GB of storage capacity).

A factor that played a critical role in enabling more data to be stored in less space, is the data compression techniques that continued to evolve over several decades, alongside advancements in computing and data storage. Starting with the use of various coding techniques over the years, all the way to AI and machine learning used today, innovative compression techniques have played a crucial role in managing the ever-growing volumes of digital data.

So, while the storage capacity has expanded massively, the physical size of data storage has considerably diminished.

Now let's look at how this phenomenal growth of data has happened across each of the 3V dimensions in detail.

Volume

The volume of data can be defined as all the available data that is out there that needs to be assessed for relevance, or what IDC defines as the global datasphere—the quantification of the amount of data created, captured and replicated across the world.²⁷ In simple terms, it is all the data that is available in the world. Every word that I am writing in this book, every text that you send to your friend or family, every image, video or post that you create online, every online transaction you make, all become part of the global datasphere. And as more and more people get connected, these interconnections create more and more data. As I discussed earlier in this chapter, most of the data has been generated in recent years. At the dawn of 2020, the global datasphere was estimated to be 44 ZB. But in 2020 alone, 64.2 ZB of data was created or replicated, defying all estimates, despite the systemic downward pressure due to Covid-19 pandemic on many aspects of our lives and businesses. In 2021, this number grew even bigger, 79 ZB, expected to reach a massive 181 ZB by 2025, in just the next three years.²⁸ What's noteworthy is that the volume of data being generated is not growing in a linear manner, it's growing exponentially. Fun fact—if a single person tried to download all the data from the internet today, it would take them around 3 million years.²⁹

The biggest contributor to this growth is the internet. As of January 2022, there were about 99,000 Google searches every second, roughly 8.5 billion searches every day.³⁰ People watched 2 billion videos on X (formerly Twitter) daily.³¹ Also, it is estimated that more than 700 million hours of videos are watched globally by Netflix users every day.³² But the biggest shift in data generation is that people are no longer just consuming content, they are also *creating it*. Consider this. Facebook has almost 3 billion users, YouTube 2.5 billion users, Instagram 2 billion users and X more than 600 million users (October 2023 figures).³³ Every day, these users contribute billions of images, posts, videos, tweets, etc. As per 2023

statistics, around 1 billion stories are posted every day and 350 billion photos have already been uploaded on Facebook.³⁴ And in 2022, Instagram users posted about 3,47,222 stories and around 650 million tweets every day. ³⁵

IoT as a volume booster

In addition to the data generated by humans, there is massive amounts of data generated by machines as well. Forty per cent of internet data in 2020 was machine-generated. IoT-generated data is growing rapidly and does not show any signs of slowing down in the future. In 2019, around 13.6 ZB of data was created by IoT devices. This number is expected to grow by around 6x, to reach 79.4 ZB by 2025. ³⁶

From close to half a billion IoT devices in 2003, to around 53 billion in 2023, IoT devices have grown by 100 times.³⁷ Healthcare has been the fastest adopter of IoT. Medical IoT solutions like blood glucose and heart rate monitoring, pacemakers, fall detection, geofencing and location monitoring are rapidly growing in number. The value of this sector—called the Internet of Medical Things (IoMT)—is predicted to reach \$176 billion by 2026.³⁸

The volume at which data is generated will continue to grow at an exponential pace as more and more digitization happens. The pandemic has further propelled this pace as more industries worldwide are forced to accelerate their digitization efforts. The digital economy already makes up for 15 per cent of the global GDP, registering 2.5 times faster growth over the physical world GDP, over the past ten years. And by 2030, it is projected to double in proportion, making up for 30 per cent of the global GDP.³⁹ Organizations are also accelerating their digitization efforts to become more agile and resilient to the ever-changing environment that they operate in.

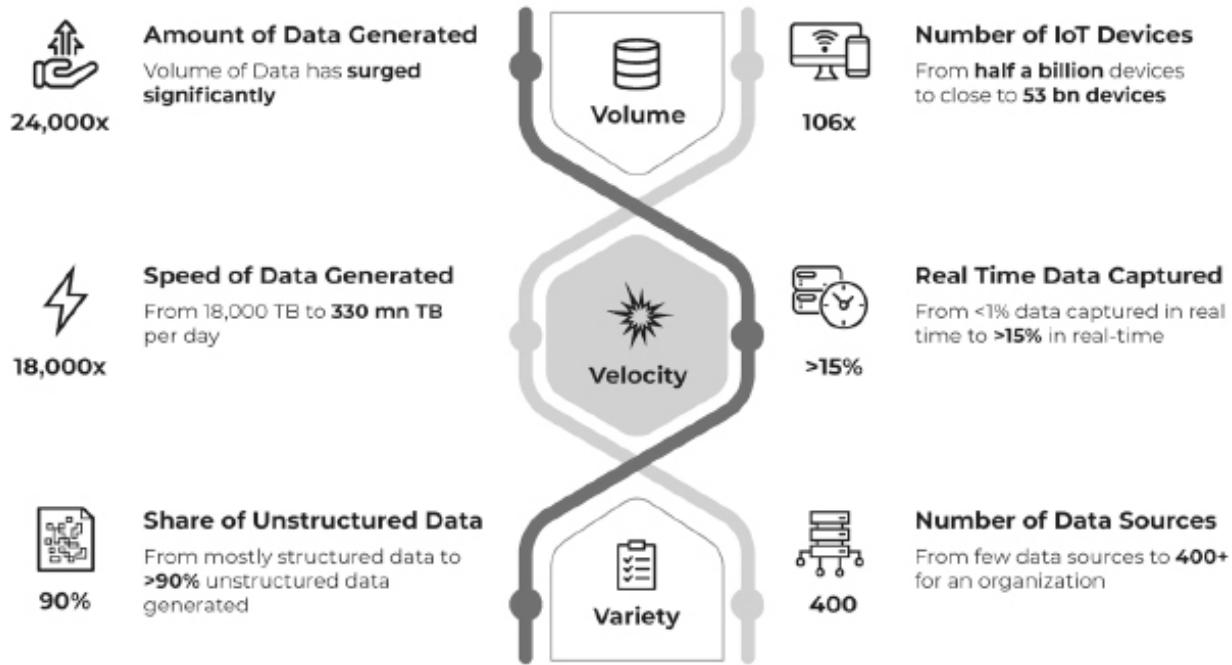
Velocity

Velocity of data can be defined as the speed with which data is being generated. In 2003, global data generated per day was 18,000 TB, but in 2023 it is 328.77 million TB per day,⁴⁰ witnessing an *18,000 times* growth, over the course of just twenty years. Staying with our internet and social media example, according to 2022 statistics, 87 million WhatsApp messages are sent every minute,⁴¹ over 500 hours of new content is uploaded to YouTube every minute, on average 6000 tweets are sent every second and around 4 million likes generated on Facebook every minute.⁴² To add to this, businesses are not only receiving data faster than they can make sense of it but are also generating data at a lightning rate. The New York Stock Exchange captures 1 TB of trade information during each trading session.⁴³ A company, on average, is estimated to generate around 23.14 TB of data every second (2000 PB per day).⁴⁴ Amazon ships approximately 1.6 million packages a day. That comes down to more than 66,000 orders per hour and 18.5 orders per second.⁴⁵

The 'real-time' of data generation

In 2003, barely any data was captured in real time. By 2025, around 30 per cent of the total data is expected to be captured in real-time.⁴⁶ A recent survey found that 71 per cent of the over 500 technology leaders they questioned believe that they can directly attribute revenue growth to use of real-time data in their organizations.⁴⁷ Mastercard leverages real-time data to analyse millions of transactions taking place on its network every day to determine which ones are likely to be fraudulent. Many of the e-commerce websites use real-time data to provide product recommendations and assist in the customer journey in real time. Companies across multiple sectors, whether it is financial services, consumer goods, automotive industry, are leveraging customer data in real time to deliver personalized products and services.

Big Data has brought dramatic changes in Volume, Velocity and Variety of data



So, as more and more companies lean on real-time data to drive value, the speed at which data will be generated and captured in the future will also continue to grow rapidly.

Variety

The third and equally important dimension on which data has exploded is the 'variety'. Variety of data can be defined as all the different types of data—structured or unstructured—that are created via a growing number of sources available today. The largest chunk, more than 80 per cent of the data, is generated in unstructured format, like emails, voicemails, handwritten notes, photos, videos, tweets, sensor data, etc. Clearly, unstructured data is what has contributed to the explosion in variety of data and will in fact comprise the vast majority of the total data in the world. In addition to that, the data has exploded in granularity owing to the fact that the number of attributes captured for each data point has also grown exponentially. For example, a weather forecasting company traditionally would capture a few metrics like temperature, humidity,

wind and precipitation. But with the technology advancements now, they are able to capture detailed metrics such as temperature at different altitudes, humidity at different levels, different types of precipitation like snow, sleet, rain, air pressure at multiple levels, atmospheric composition—oxygen, nitrogen and so many more. While it affects the volume of data, more importantly it adds to the number of variables being captured. Even with structured data, although it makes up for a mere 20 per cent of all the data

generated,⁴⁸ owing to the increasing number of attributes being captured, the complexity around variety holds true here as well. For example, for a credit card customer, a company captures multiple attributes such as, personal information like name, address, phone number, etc; identification information like social security number (or equivalent identification number) and date of birth; financial information like credit card number, credit limit, current balance, transaction history, payment history and so on.

And the more varied the data is, the more effort it takes to harmonize it for consumption (for example, creating insights).

Data sources go from zero to hundreds

Data sources are unbounded today. Anything today can be a data source. Smartphones, computers, websites, social media networks, e-commerce platforms, IoT sensors, wearable devices ... the list goes on. A survey of more than 200 executives of North American organizations with at least 1000 employees reveals that there are more than 400 data sources available today per organization. And about 20 per cent of organizations are leveraging more than 1000 different data sources for analytics and drawing insights.⁴⁹ Now imagine leveraging 1000 different types of data sets that have different format, varying speed and varied levels of complexity. Consequently, traditional data processing tools fall short on effectively analysing such varied and complex data sets and require more sophisticated tools and more specialized skills to translate these into standard and comparable formats for effective analysis.

The additional dimensions of Big Data: +3V

Over the years, several attempts have been made to refine this 3V model, and some additional dimensions have been added to it.

Back in 2011, IBM added the fourth dimension to Big Data—**'Veracity'**. Veracity refers to the extent of reliability, consistency, accuracy and trustworthiness of the data collected. The data that is being collected by the organization comes in from different sources and at different times and at varying speed. Therefore, there is a high chance that this data could be incomplete or inaccurate or may not be substantial enough or relevant enough to provide accurate, real or valuable insights. Veracity is the level of trust in the data that is being collected.

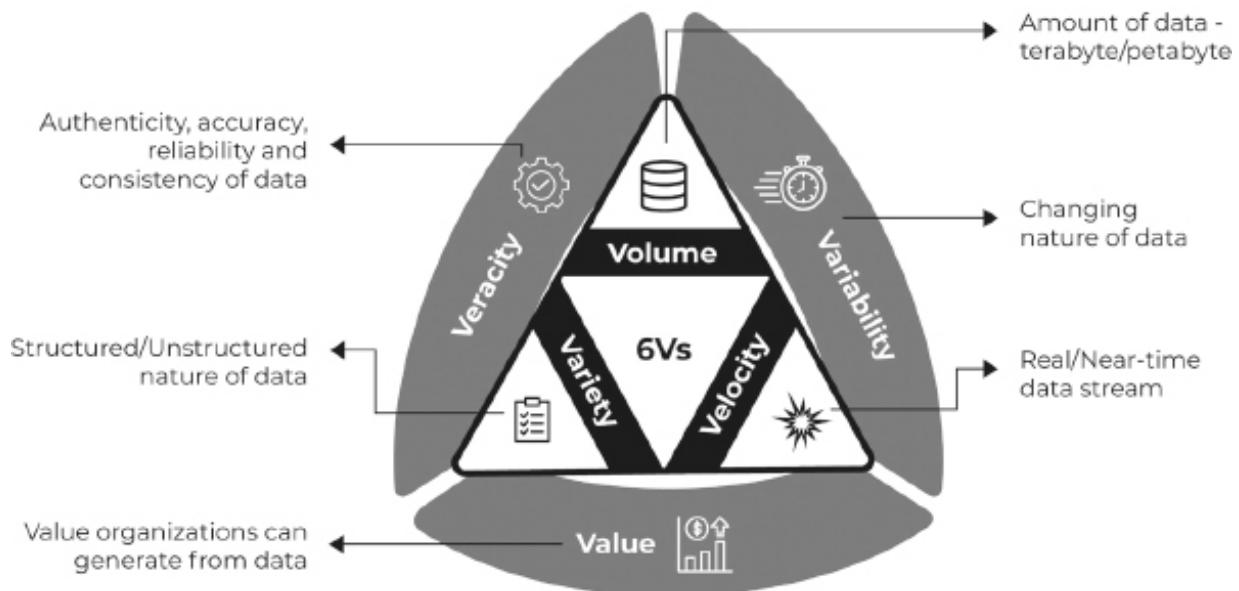
Then, down the years, as the nature and complexity of data kept evolving, making it more dynamic in nature, another new dimension was added—**'Variability'**. Variability is the changing nature of the data to be captured, managed and analysed. In today's dynamic environment it is naïve to expect that the meaning of data would remain static. Also, newer dimensions are added every day while some dimensions become obsolete with time. Therefore, if the meaning and importance of data keeps changing constantly, the sanctity of that data could be in jeopardy.

And another dimension that got added somewhere along the way was '**Value**'. Value is the end game, the outcome or the impact. It is the direct measure of what organizations can achieve using the data gathered and processed. It is the process of extracting knowledge and insights from all the structured and unstructured data, without loss or distortion of the real meaning of that data. Big data scientists consider this as an important aspect, since the valuable insights lie hidden somewhere within all the data which needs to be extracted at the right time for the right reasons.

Over the years, many such dimensions have been added to define the characteristics of data. We have the 4Vs, the 5Vs and even the 7Vs of data out there. But I believe that core 3Vs that were initially coined by Doug Laney, continue to be the core dimensions to understand data revolution. So, for the purpose of this book, I will

continue to focus on these three dimensions only—volume, velocity and variety.

While the definition of Big Data has expanded, Volume, Velocity and Variety still remain the core dimensions



In my opinion, veracity and variability are aspects of data quality, which I will discuss in detail in Chapter 12, Data Quality, and value—which is the outcome or impact will be delved into in the next chapter.

Key takeaways

- The sheer magnitude at which data has grown over the past twenty-four years is an unparalleled phenomenon in human history.
- The dramatic growth and adoption of digital technologies have resulted in an exponential increase in the capacity to generate, capture and store data, thus giving rise to the phenomenon known as 'Big Data'.
- While structured data exhibited notable growth, the surge of Big Data is primarily owed to the proliferation of

unstructured data.

- Data explosion is seen across three core dimensions—the 3Vs: volume which refers to all the available data at any given moment, variety that encompasses the diversity in data generated, and velocity that characterizes the rapid pace at which data is being generated.

Data, the Fuel for the Digital Age

Impacting Every Aspect of Life and Business

'Data really powers everything that we do.'

—Jeff Weiner,
Executive chairman, LinkedIn

Data is ubiquitous. Every time we interact with the digital world, we consume data and we generate it at the same time. Owing to the phenomenal growth of data, it has exploded in all directions, it has become an integral part of everything we do and how we do it. In this chapter, I want to emphasize the magnitude of change that data has brought about in every aspect of human life, creating unparalleled opportunities for businesses to generate value like never before.

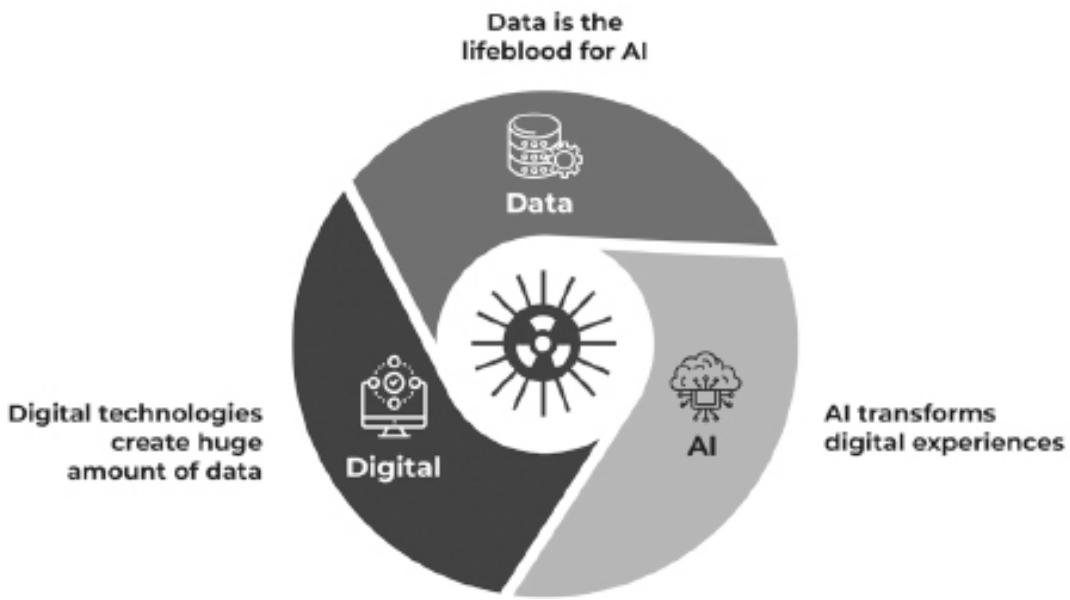
The wheels of change are turning faster with data

Data, in fact Big Data, today plays such a prominent role in our lives that it is impossible to summarize all the ways in which data is powering the world. It has truly become a significant force that has been shaping the world that we live in. This transformative power originates from the interplay of data, digital and AI. Data is the lifeblood of digital technologies, allowing them to achieve the unthinkable. Digital technologies are in turn generating more data that is continuously enriching the data pool. This data then becomes the fuel for AI including generative AI, allowing it to learn, adapt and transform digital experiences in unprecedented ways. This, in turn,

generates even more data, further enriching the data pool and fuelling the cycle.

Data, digital and AI are tightly linked in a powerful virtuous cycle. As more data becomes available, AI becomes more powerful, leading to better experiences delivered by digital technologies and even more data generation. This feedback loop fuels explosive growth and unlocks incredible possibilities.

The virtuous cycle of 'Data + AI + Digital' propels an explosive transformation, with each element seamlessly feeding into and enhancing the others



Since most of the data has been generated in recent years and the amount of data collected every day is mind-boggling, it has brought with it immense potential and capability which humans have access to for the first time in history but are still grappling to fully understand. Experts, data scientists, data gurus all over the world are working to apply the insights gained from data in numerous ways. In fact, data that is now available openly and abundantly has opened opportunities for the common man and the entrepreneur alike, to reimagine using data in highly unique ways. And within a very short period, quite unassumingly, data has transformed every aspect of how we live, interact and conduct our businesses.

The pandemic accelerated the shift towards digital adoption as organizations and individuals worldwide had to swiftly embrace digital tools for survival and operational continuity. This transition became an imperative rather than an option, as across so many aspects of life and for so many industries there was a sudden shift from in-person to virtual interactions. While this isn't the first pandemic the world has witnessed, widespread adoption of digital technologies enabled unprecedented data availability and made it easier to adapt to virtual business models, despite widespread disruptions. As a result, changes like hybrid work, online education, virtual healthcare and mental health resources have become the 'new normal'. Needless to say, this shift is further adding to the global datasphere, making it richer and more comprehensive by the day.

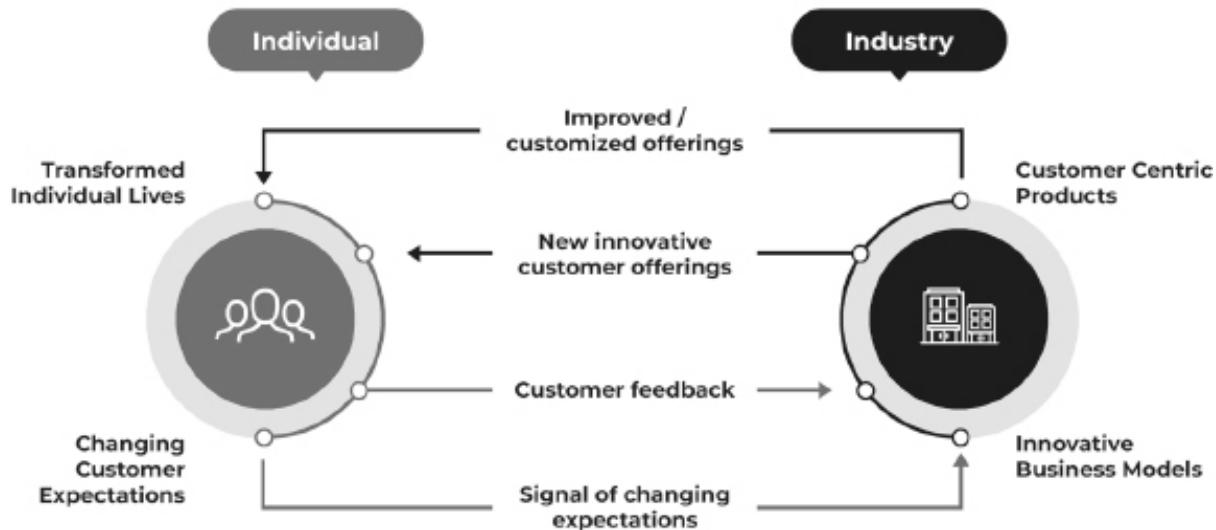
The world of data is powering the wheels of change all around us. Individual lives have been revolutionized by data, so much so that the traditional ways seem distant now. Data is shaping daily routines, from morning alarms set by AI assistants to consuming news and weather updates and even bedtime reading on e-readers. And as consumers engage deeply with data, organizations seize opportunities to craft unique offerings. Organizations have infinite opportunities to tap into all the data available out there, using advanced digital technologies to innovate customer-centric products, optimize operations and expand capabilities. Additionally, this era is also witnessing the mushrooming of many innovative start-ups and disruptive business models who are leveraging data like never before, who have redefined and are continuously pushing the bar in terms of customer expectations. They come in and completely change the very rules of the game, in fact, at times they change the game altogether.

This interplay between individuals and organizations that has been creating an unending loop of change enabled by data, is picking up pace in the data-first world and setting the wheels of change to turn faster all around us.

As I said before, it is literally impossible to summarize all the ways data has helped redefine the world around us but let me give you a

sneak peek into a few of the myriad ways data has done so. Let's start with how data has changed the way we live.

Big Data has transformed the interplay between individuals and industries



Our way of living revolutionized

Data all around us has transformed every aspect of our lives. We can argue that it is not always for better, and we will get to that debate later on in the book. But the reality is that we can no longer imagine life without digital technologies, powered by data. And this trend has been further intensified due to the challenges and restrictions posed by the Covid-19 pandemic. Let us look at these shifts across some select areas.

Personal guide leading the way

Gone are the days of asking for directions; GPS has transformed the way we travel. Today, you simply input the destination and GPS handles the rest. It stores all your journey data—frequent routes, favourite spots and important addresses—accessible at our convenience. Before any trip, you can comprehensively plan from start to finish, including 'where to go', the route, duration and post-arrival activities, all at your fingertips. Platforms like Google and

Apple maps make it super easy to navigate routes with turn-by-turn guidance and congestion updates in real time. This reliance deepens as familiarity grows, so much so that globally, about one in every two individuals say that they would not be able to reach their destinations without a GPS.¹ Amid the pandemic, GPS served beyond navigation, enabling monitoring of Covid-19 spread. Countries like Israel employed mobile phones for contact tracing, monitoring those near an infected person within 15 minutes of diagnosis.² India's Aarogya Setu app similarly tracked routine interactions, using GPS, to trace contacts in case of infection.

Entertainment on demand

Entertainment has been transformed remarkably as well. From waiting for days for a new episode of your favourite series or standing in long queues to catch your favourite movies, we are now in the age of any-time, anywhere, on-demand entertainment with streaming services. It has completely transformed our expectations and the way we consume content. We are attuned to receiving content recommendations, promotions and services that align more closely with our individual tastes and needs and are available to us in real time and on demand. And as we get accustomed to on-demand entertainment, our expectation for an increasing abundance of readily accessible content also escalates. For example, today the concept of single episode release per week has become near obsolete. We can now 'binge-watch' the entire season of a series in one go, or at our own convenience. In fact, as per 2020 US survey, around 70 per cent of viewers between the ages of eighteen and forty-four binge-watch TV shows and movies regularly.³ During the Covid-19 lockdown when public entertainment had come to a standstill, over-the-top (OTT) platforms kept us entertained. The global OTT market anticipates a 19 per cent compound annual growth rate (CAGR) from 2019 to 2026,⁴ driven by elevated demand for OTT services and gaming during the pandemic, a trend that is here to stay.

Prevention is better than cure

Healthcare is another sector that has been significantly revolutionized by data and digital. While this sector has historically been a pioneer in adoption of digital technologies, a notable recent transformation is the shift towards prevention over cure. Consumers who are becoming increasingly health conscious are well informed and are actively managing their well-being. From proactively tracking their health to finding the best course of treatment, consumers are turning to digital devices and online resources for guidance. Wearables play a pivotal role, capturing real-time health data and empowering individuals to monitor their health closely. The number of wearable devices almost doubled in the span of five years, from 593 million in 2018 to 1.1 billion in 2022.⁵ The pandemic further accelerated the shift to digital, with patients relying on tele-medicine for consultations and treatment. Tele-health usage surged during the pandemic, reaching seventy-eight times the February 2020 level in April 2020. Although it has stabilized at thirty-eight times post-pandemic in 2021, online consultation's lasting presence is evident.⁶

Socializing goes digital

Social media, an aspect born from digital, detailed earlier in Chapter 1, Data Explosion, significantly altered the way we socialize, connect and communicate. It erased physical boundaries, uniting people in unprecedented ways. The entire Globe has shrunk down to a dot, in the [dot.com](#) era. The Republic of Facebook is now the world's largest country by population, with 2.99 billion.⁷ Virtual inhabitants, more than double of world's most populated country India, with a population 1.43 billion (as of July 2023).⁸ Housing such a huge community of virtual inhabitants, social media has become a leading generator of content, constantly capturing user data, moulding our lifestyles and our perspectives as well. Beyond a source of learning and sharing, it has the power to create awareness, amplifying voices and fostering unity at a global level. For instance, the

#blacklivesmatter movement swiftly transcended social media, altering opinions for 23 per cent of US users.⁹ Amid the pandemic, social media became a lifeline for those confined at home, enabling continued interaction, sharing and connection. It served as the primary means to stay relevant, connect with loved ones and seek information and assistance. At its height, social media countered isolation and boredom, offering remedy for loneliness. Globally, July 2020 witnessed a 10.5 per cent surge in social media usage compared to July 2019.¹⁰

Virtualization of shopping

What can ever be better than being able to buy products from the comfort of your home, sitting on your couch. The days of store hopping, haggling and travelling from one shop to another to find the right item, have been replaced by the ease of online purchases. Today we have multiple products from multiple brands available on one platform, and these can be delivered to our doorstep at the click of a few buttons. From kittens to cars, literally everything is available for sale online. In 2022, retail e-commerce sales were estimated to exceed US\$5.7 trillion worldwide and is expected to account for 23 per cent of total retail sales by 2025.¹¹

E-commerce has broken down geographical barriers, enabling customers to access products and services from around the world. They can explore a diverse range of offerings that might not be available locally. Tailored product recommendations, based on their individual preferences, purchase history and browsing behaviour, ensure a more personalized buying experience. Furthermore, online platforms allow easy comparison of prices and features across different brands and retailers, which helps customers find the best deals and value for money. Furthermore, diversified shopping experience with features like virtual try-ons, interactive product displays and augmented reality has enabled shoppers to enjoy in-store buying experiences online as well. Flexible delivery options are also available, like same day and next day delivery and even

subscription-based services, to accommodate varied delivery preferences and urgency levels.

The pandemic further revolutionized this trend as customers were unable to visit their stores in person. They relied on online platforms to get everything from necessity to luxury items. The pandemic added an additional 19 per cent in 2020 and 22 per cent in 2021, to the regular forecasted sales growth rates for those years.¹²

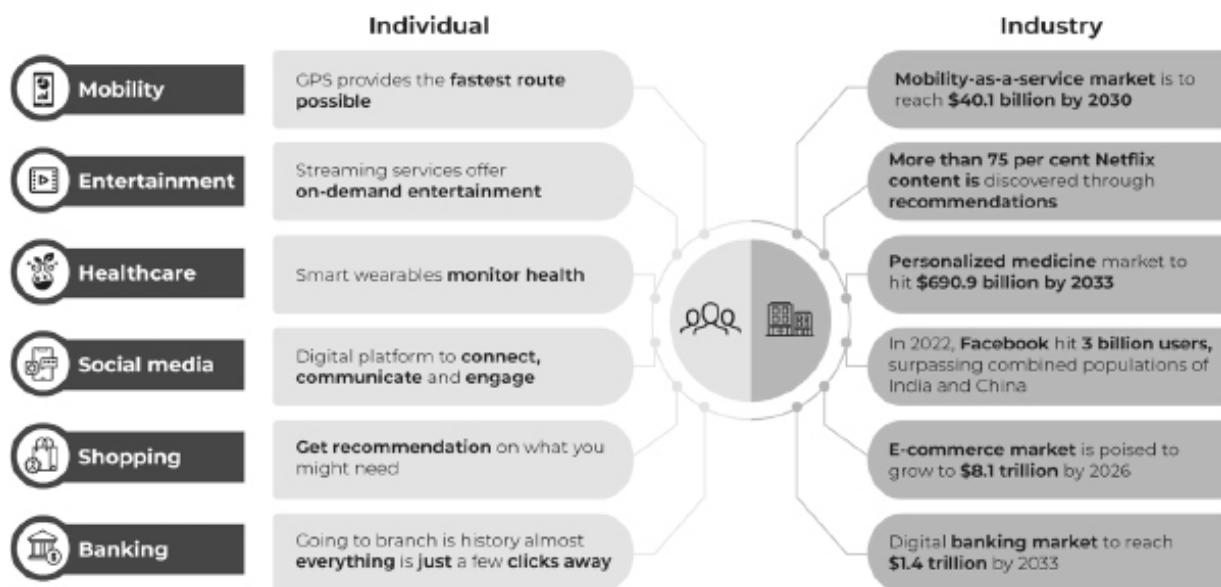
Faceless, cashless transactions

Though it feels like another lifetime, not so long ago opening a bank account demanded a time-consuming branch visit, similar to navigating a maximum-security prison with every move being monitored by the bank's security guard. The process involved long waits, tedious form-filling, getting subjected to managerial scrutiny and a background check that took days, if not weeks. Basic transactions like withdrawing cash required scheduling your time to match the bank's operating hours, and then waiting some more. Today, banking is largely driven by customer experience. Tech-savvy customers are opting for seamless digital channels, handling bill payments, fund transfers, investments and various other banking and financial activities with a click. We are offered more customized and comprehensive financial solutions, tailored to suit our financial needs and aspirations. Simplified banking processes like e-KYC using GPS-based verification, expedite loans from hours to days. Mobile wallets, digital-only banks and personalized services clearly represent the evolving banking behaviours and expectations shaped by Big Data.

The pandemic turned us all into digital-first users, propelling a swift adoption of digital banking as financial transactions moved online, with two out of every three of adults engaging in online payments, globally.¹³ In fact, a June 2020 survey revealed that 44 per cent of eighteen- to thirty-four-year-olds adopted online/mobile banking for the first time during the Covid-19 crisis.¹⁴

As a result of this seminal shift in consumer behaviour and customer expectation, organizations have had to evolve their business models to adapt and deliver value to their customers. Not only that, new and innovative customer offerings have also been rolled out to cater to the tech-savvy, rapidly evolving customer base.

Data Revolution has completely transformed human life and created tremendous opportunities for businesses



Unprecedented value generation opportunity for industries

As the way of life continues to evolve in the data-first world, newer opportunities emerge for businesses to add value. Access to abundant data throughout the value chain enables companies to enhance products and services. Every company, regardless of the business they are in, has become a technology company, leveraging data to unlock hidden value from unexpected avenues and innovate much more rapidly and freely.

As a result, all industries alike have witnessed dramatic changes in the customer and business dynamics over the past few years, leading to a tectonic shift in the business models and the competitive landscape. As I also highlight in my previous book,

Winning in the Digital Age, digital disruption has triggered some common shifts, transforming every business into a consumer-centric one. Disintermediating to reach customers directly is redefining the value chains across industries, and new competitors with more innovative business models are further redefining industries. Consumers have wider choices as newer products are being introduced in the market every other day, putting downward pressure on prices, and both customers and regulators are demanding more transparency. All these shifts are enabled by data empowering companies to go beyond their core capabilities to innovate and deliver enhanced value to their customers as well as their stakeholders.

Let us look at how data has redefined the rules of the game for a few prominent industries.

Banking: Clicking away financial worries

Up until a few years ago, every financial transaction was highly controlled and complex. But with the advent of online banking, came the benefits of ease of transactions at lower transaction costs, easier integration of services and more targeted marketing capabilities. The adoption of mobile banking in the early 2000s enabled financial management from virtually anywhere, any time. In 2022, around 2.2 billion individuals have used digital banking services.¹⁵

The trend of digitization started when leaders of financial services firms realized that the majority of their users were available on digital channels. As most of them were transacting via mobile or websites, omnichannel came into play. Customers' transaction and activity data also opened opportunities to understand their behaviour and financial journey better. Fintech companies and digital-only banks leveraged this data to create unique and comprehensive offerings for their customers, making financial services easily accessible at substantially lower costs. Today, mobile banking, Unified Payments Interface (UPI), online bill payment, peer-to-peer lending and transactions and online credit have become the norm.

The share of market capitalization of conventional banks and financial institutions slipped gradually, from 96 per cent in 2010 to around 72 per cent in just a decade.¹⁶ Fintech firms like PayPal and non-bank payment firms like Visa are chipping away at their share of the market capitalization. Financial Services giant Goldman Sachs foresaw the digital shift and its transformative potential early on. In 2019, they recruited a chief information officer (CIO) with years of experience at Amazon, who emphasized a customer-centric approach to building software—based on insights from customer data. Considering data as the lifeblood of any strategy, they invested significantly in data quality, lineage and platforms to establish a robust foundation.¹⁷

Fuelled by Big Data, emerging digital trends enabled by data and analytics are reshaping the industry and shifting customer expectations. Credit approvals now take minutes, through consolidated credit history and e-KYC. The FinTech Lending Market foresees 27.4 per cent CAGR growth, reaching \$4.96 trillion by 2030.¹⁸ Real-time fraud detection, facilitated by machine learning (ML) models, combats money-laundering proactively, projected to grow at a 22.6 per cent CAGR to \$182 billion by 2029.¹⁹ Digital payments surged, driven by analysis of data generated at multiple touchpoints, set to reach \$361.30 billion by 2030, from \$88.1 billion in 2021, a 20.5 per cent CAGR.²⁰

Travel: From ownership to usership

Born during a snowstorm on the streets of Paris in 2009, the idea of being able to hail a cab using a mobile phone revolutionized the concept of car ownership forever. Uber, originally named UberCab, used geophysical location (GPS) and payment data to make affordable rides instantly available by clicking a few buttons on the mobile phone. And the concept of shared mobility was born. Shared mobility revenue is forecasted at \$1.53 trillion in 2023, rising to

\$1.79 trillion by 2027. Shared vehicles are set to reach 5.1 billion by 2027.²¹

E-hailing—a service provided to book public transport services through electronic applications, dynamic shuttle pooling—like door-to-door detours but combining close-by stops, peer-to-peer car sharing—privately owned vehicles shared with others, car rentals and shared-micromobility—shared e-scooters or bicycles, are all gaining momentum as more consumers strive to become asset-light and environment-conscious. And all this has been made possible by Big Data and analytics through navigation data and real-time insight generation for demand estimation and price optimization. The industry is gradually but noticeably shifting towards usership from ownership across geographies, developed to developing economies alike.

The original equipment manufacturers (OEMs), or automotive manufacturers, who have been hesitant to invest in this business model, owing to the fear of cannibalization of vehicle sales, require a mindset shift from selling vehicles to providing mobility services. But as the global car sales continue to show timid growth, OEMs are finding new ways to capture a piece of this market too. Taking cues from the likes of Zoom cars, major luxury car companies like BMW, Porsche, Lexus, Audi, Land Rover and Mercedes-Benz, have all started testing subscription programmes—a programme that allows drivers to pay a monthly fee or rent to use the vehicle. The global mobility-as-a-service market was estimated at \$ 5.7 billion in 2023, expected to reach \$40.1 billion by 2030, growing at a CAGR of 32.2 per cent during that period.²²

Shopping: Amazonification

Yes, 'Amazonification' is a widely accepted word now! Starting with selling books to customers over dial-up modems, Amazon has come a long way to become one of the most valuable companies of the world today. And it has revolutionized the way we shop forever. In fact, it revolutionized customer expectation to such levels that over

the years, customers have started expecting similar experiences from other industries as well. That's some feat to achieve.

With their eyes set on becoming the 'everything store', Amazon leveraged customer data in the most innovative ways, making some monumental shifts in customer offerings that shifted the retail landscape from offline, in-store shopping to online, at-your-doorstep delivery business. Amazon owns almost half of the US retail e-commerce market,²³ and is the largest e-commerce player globally.

Amazon has mastered the art of leveraging customer data and using artificial intelligence and analytics to better understand and predict their buying behaviour and the shopping journey and worked towards making the shopping experience easy, frictionless and cost-effective. They pioneered in providing multiple innovative customer experiences with personalized recommendation engines, one-click ordering, package pickup at Amazon hubs and lockers, real-time tracking and tracing, easy returns and replacements, Amazon Dash for ordering products with the single touch and in-home delivery with Amazon Key.

The Amazon business model became a threat to brick-and-mortar retail in more ways than one. In fact, as e-commerce surged, questions arose about the future of brick-and-mortar stores and local shops. A prominent example of that is when they significantly impacted the popularity of Black Friday schemes in stores and continue to do so. In 2023, online sales grew 8.5 per cent on Black Friday whereas overall retail sales increased only 2.5 per cent.²⁴ Amazon's online version of Black Friday liberated shoppers from long queues, frantic searches and the anxiety of losing coveted items. Owing to this Amazonification, High/Main Street retailers had to significantly transform their business models and forge connected value chains to offer omnichannel experiences. Retail giants like Walmart decided to rethink their strategy and go online in the year 2000. Today, all retail stores try to reach out to shoppers anywhere, any time and through multiple channels and devices. Even the mom-and-pop stores have evolved to facilitate doorstep delivery within minutes. With every one out of five purchases being made online,

the global e-commerce market is estimated to reach \$6.3 trillion in sales by January 2023, poised to grow to \$8.1 trillion by 2026.²⁵

Entertainment: Netflixization effect

Netflix, which started its journey as a humble DVD distributor, has grown to become a truly disruptive force in the television and movie industry. It did that on the back of streaming service, using the internet to deliver content right to the viewers on demand, playing a significant role in popularizing the concept of OTT—a direct-to-consumer video content platform. The key to them building a transformational business model and garnering such high viewership lay in the use of data and analytics. Using data on each individual, it created a detailed viewer profile. One of the many things it used data for was personalized recommendations for movies and TV shows based on its subscribers' preferences. According to Netflix, 80 per cent of watched content is based on personalized recommendations.²⁶ It also collected customer interaction and feedback data on TV shows, like the duration a customer viewed, the pauses they took and whether they resumed watching it, abandoned shows, the device used and many such data points to feed into its recommendation algorithm.

Netflix also leveraged insights to craft curated content. Original content was produced, tailoring content for diverse viewers based on its understanding of preferences of varied viewer demographics, regions and ethnicities. Through innovative use of data and analytics, it identified underserved demographics and harnessed fresh talent. For example, it created content focusing on neglected segments like young women aged fourteen to thirty-four. It also used data and analytics for targeted marketing, customizing the trailers for the same TV show for different audiences and their viewing preferences.

The revolutionary shift towards streaming content has led to a substantial loss of viewership for many top-tier cable networks. For example, if you look at the US cable network market, between 2011

and 2021, Disney channel saw a 90 per cent drop in viewership, ABC Family saw a 71 per cent decline, Fox Cable lost 67 per cent of its primetime audience and AMC network saw a 57 per cent decline.²⁷ Soon enough, giant TV and cable names like NBC, HBO, ABC, CBS all had to evolve their traditional delivery networks to deliver content online, anywhere, on demand. Most have now launched their own streaming services, have incorporated recommendation engines and many are also now generating original content to compete with the likes of Netflix and Amazon Prime, and attract more viewership.

Advertising: Shift to social media

The newspaper industry saw a significant decline in ad revenues that migrated to digital media. As a result, over the past six years leading up to 2023, the global print advertising market has undergone a significant reduction, projected to decrease by half to reach \$47.2 billion this year.²⁸ Furthermore, the OTT platforms boasting of providing ad-free access to entertainment meant a big hit for brands spending millions of dollars on TV ads in the hope to reach their target audiences on a wide scale. The ad spend on digital was estimated to account for 55.5 per cent of the total global ad spend, double of the TV ad spending in the year 2022, estimated at 26.1 per cent.²⁹ This TV ad share is expected to gradually decline further in the coming years while digital spend will continue to rise. Now, as OTT platforms increasingly consider implementing a hybrid, ad-supported and subscription-based model for financial sustainability, the ad spends on TV would decline even more.

So, besides television, social media, with over 4.6 billion users in 2021, expected to reach 6 billion by 2027,³⁰ has become the most preferred channel for advertisers today. By leveraging diverse data sources, advertisers can hyper-target audiences with personalized content based on demographics and user behaviour. Social media appeals to advertisers as it facilitates improved user engagement and conversions, cost-effectiveness in reaching wider audiences and swift growth in brand awareness. Sharing and re-sharing capabilities

on social media amplify the reach, surpassing any traditional advertising method.

Social intelligence tools capture insights from various online platforms, aiding consumer profiling, trend analysis and behaviour understanding for creating highly targeted campaigns. Additionally, brands can continuously analyse social media conversations, sentiments and feedback to refine products and offerings, making it the most effective channel, both in terms of cost and impact. Organizations can reach their customers quickly and at scale, enabling customers to easily interact with the brand and become part of the brand community.

Healthcare: Personalized treatment and care

The healthcare industry, with its intricate ecosystem involving hospitals, doctors, pharmaceutical companies, insurance providers and patients, has historically lagged in digitization due to its complexity. Therefore, medical treatments have traditionally focused on standardized care. But in recent years there has been mounting evidence of patients responding differently to the same treatment owing to multiple factors, of which their biological makeup is the most critical. Therefore today, personalized medicine aims to tailor treatments based on unique disease characteristics specific to each patient or patient subgroup. The personalized medicine market, valued at \$326.7 billion in 2022, is projected to grow at a 7.8 per cent CAGR from 2023 to 2033, reaching \$690.9 billion.³¹

For example, cancer treatment emphasizes the role of genomics in tailoring unique treatments based on individual genetic information. Each patient's unique genetic makeup, coupled with environmental factors, contributes to the way cancer develops. Analysing these elements requires analysing extensive data, including an individual's estimated 19,000–20,000 genes, environmental influences and socio-economic conditions.³² It is a staggeringly complex exercise that requires integrating enormous data sets and analysing them using natural language processing, advanced AI and ML, to derive

actionable insights. This approach empowers clinicians and scientists to devise personalized treatment strategies for each of their patients.

Beyond personalized treatments, data is also revolutionizing healthcare delivery. Unlike the past, where capturing comprehensive patient history and progress was challenging, access to diverse technologies today enables comprehensive data collection. Electronic health records (EHR), remote patient monitoring tools, wearables and health apps provide a wealth of information. Frequent monitoring aids early disease detection and continuous care, extending beyond hospital walls. This shift to continuous care outside hospitals has significantly transformed healthcare delivery, resulting in improved patient outcomes.

In addition to this, with all these technologies available today, the amount and variety of data being generated and captured is growing day by day. And this accumulating individual data is expanding the pool of information available about the population, leading to inclusive and balanced representation in trial samples as well. Which in turn helps develop more effective treatment and cure options resulting in better patient outcomes.

Blurring of boundaries and expectations

As a result of continuously evolving customer expectations as well as the business and competitor landscape, organizations are now becoming more fluid, and are moving towards becoming a one-stop solution to customer requirements. Using data and technologies, they can stretch beyond their own area of expertise to integrate multiple solutions under one hood. Today a mobile service provider is not just providing connectivity solutions, but are also providing payments, shopping, gaming, entertainment and many other solutions as well.

The Amazonification of services and Netflixization of experiences has not only affected their own respective industries but has transformed cross-industry expectations. Today, personalized experiences have gone beyond the purview of the retail industry. Equipped with access to data, B2B customers are also expecting a

similar kind of personalization from chemical manufacturers. For example, the 'Lab Assistant' at BASF, is a web-based application that provides full access to data on complete product data and exclusive formulations that enable customers to find the right raw material and formulation ideas to create customized product formulations.³³ Similarly, omnichannel experiences are not just expected from entertainment providers today but also from insurance providers. Now insurance is available to be bought online, over phone or in-store. Customers not only expect the convenience of buying online, but also similar levels of efficiency and convenience as top-tier e-commerce sites. And since insurance purchases are highly personal and have many considerations, customers prefer multiple touchpoints before finalizing a purchase.

In addition to that, we have increasingly come to expect one-stop shop solutions stemming from Amazon-like business models that we have got used to. So, we expect even our financial services provider to provide us with an array of services under one roof. Today, in a world where customers have a wide range of options to choose from, they are ready to pay a premium for ease of use or access. They would prefer interacting with a single service provider who can cater to different customer requirements, like savings, investment, insurance, bill payments, loans or even shopping, and deliver desired customer experience. We have come to expect our financial partner to provide connected seamless services across all these offerings, available at our disposal through a single app or website, accessible through least number of clicks, with minimal documentation and shortest lead time. Today, a car loan or a home loan, is available pre-approved by your financial service provider, can be availed via the bank website or mobile app with just a few clicks and the money is transferred within a few hours or a day or two.

That is the amazing feat that data has enabled individuals and organizations to achieve. And like I said, this is just the beginning. It's just the tip of the iceberg. Can you imagine what we can accomplish if we are able to go deeper and unleash the full potential of data?

Key takeaways

- Data fuels digital technologies, enabling unprecedented applications. Digital technologies, in return, spawn more data, adding to the data explosion, further enabling widespread AI use. And AI is enhancing digital experiences and data itself. This recursive relation between digital, data and AI is one of the most powerful transformative forces in the world today.
- The omnipresence of data has catalysed a comprehensive transformation across every aspect of our life, so much so that we now find ourselves unable to envision life without digital technologies fuelled by data.
- The evolving lifestyles are opening massive opportunities for businesses across all industries. Abundance of customer data has enabled firms to deliver Amazon-like services and Netflix-like experiences. Organizations are now able to extract untapped value and innovate like never before.
- The interplay between individuals and industries is redefining industry landscapes, blurring industry boundaries and transforming cross-industry expectations.

3

Value Reimagined

Framework for Realizing Transformational Value from Data

'What we see depends mainly on what we are looking for.'

—Sir John Lubbock,

Leading mathematician, scientist and author

Data is the raw material available to us in abundance. It is both an art and a science to mould it in ways that it becomes a powerful tool in the data-first world. Yes, data has truly transformed our lives and opened doors to opportunities for organizations to create unparalleled value. But how does data generate such value, and how can AI help realize transformational value? This question is not well answered today. In my conversations with CXOs, while they are excited about the value of data in general terms, most of them do not have specific ideas or plans to realize value from data for their businesses. So, let's take a deeper dive into the life cycle of data management to understand the process of value creation.

It's an ocean out there

About 97 per cent of earth's water is ocean, which ideally should be enough for everyone living on this planet. But despite being surrounded by so much water, 1.1 billion people worldwide do not have access to water and about 2.7 billion people face water scarcity for at least one month of a year.¹ Why? Because ocean water is not fit for use in its original form and the 3 per cent freshwater just isn't

enough. So, we are stuck in a situation aptly surmised by the famous lines 'water, water everywhere and nor any drop to drink'.

By the way, have you ever tried to drink ocean water? If you do, it will make your blood salty which would require your kidneys to work overtime to flush out all that salt. And to do that it would require even more water. And thus, not only does it not quench your thirst, but it also ends up making you thirstier. But there are some marine animals that survive on ocean water. It is because they have super-efficient systems to process it and some seabirds have special glands to remove the salt.

That is exactly the problem with data too.

We all know that the ocean of data we are surrounded by is full of unimaginable potential. But all this data in its basic form is useless like the ocean water and deriving value from it is becoming more and more complex and expensive. It therefore needs to be approached more systematically. Unless we are able to do so, it will not deliver the desired impact. In fact, it might lead to a tsunami-like situation where you can easily get swept away. What we need is a compass that can navigate us through these choppy waters of Big Data, to give us a solid sense of direction.

A structured approach or framework is crucial for enterprises to maximize value from data. Without which organizations can find themselves in a 'rabbit hole-like' situation, where solving a problem can create a cycle of new challenges.

But before I get into the discussion of how to generate value from data in a structured way, it is important to understand the data management value chain, the various stages and processes involved in effectively managing data within an organization. The data management value chain represents the end-to-end journey of data, encompassing a series of interconnected activities that contribute to maximizing the value of data.

The data management value chain

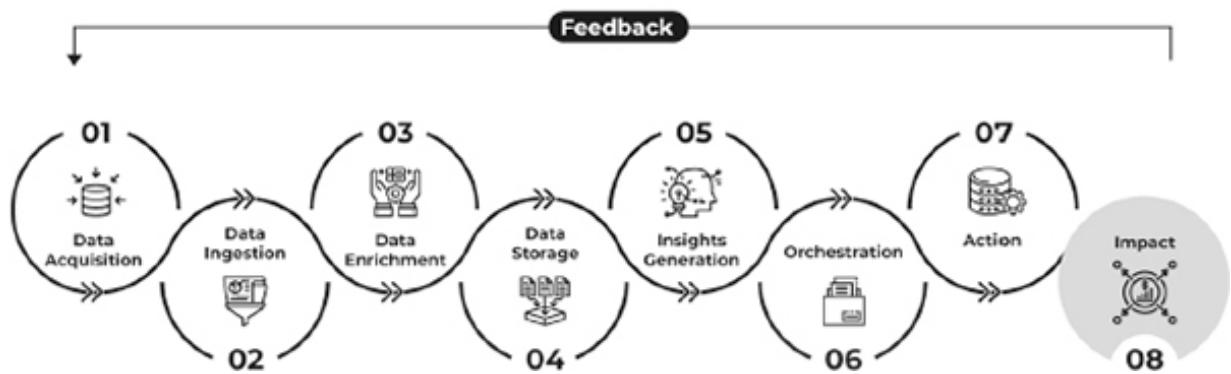
As discussed in the previous chapter, data is growing at an exponential rate and organizations are now increasingly capturing

more and more data. This overwhelming inflow of data is putting strain on existing resources of the organization. In fact, in a 2020 survey, six in ten executives acknowledged that the rate at which data is expanding exceeds their organizations' capacity to cope with it, and almost half of them said that their organizations would not be able to deal with such rapid volume growth.² So, to effectively manage such burgeoning volumes of data, it is critical for organizations to first understand the data management value chain and the process of value generation from data at each stage.

Since data is considered the oil of the twenty-first century, let's take the example of an oil-refining process to better understand value generation. Crude oil, in its original form, is not useful for anything. It is just a dark, murky liquid that we all know has tremendous value hidden in it. What it requires is a systematic process of cleaning and refining to bring out the best. Only when it goes through numerous levels of treatment, in the exact sequence and under precise conditions, does it transform into desired forms, suitable for intended use. And every stage of the refining process results in enhancement of the input material and is essential to the entire process. And as the intermediate products move up the refining chain, it gets augmented progressively. First, oil is extracted from oil wells and transported to the refinery for further processing. Once it reaches the refinery, it is fed into the distillation towers where the crude oil is heated, causing the oil to vaporize. As it cools down and descends to different heights, the oil separates into different components called fractions. Lighter fractions like gases, gasoline and kerosene rise to the top, while heavier fractions like diesel and residual fuel settle at the bottom. Each of these fractions is then further processed separately, through various conversion processes like cracking or reforming to convert heavy- or low-value fractions of crude oil into lighter, more valuable products. The various component streams are blended in the blending stage to get various grades of gasoline, which are then stored in tank farms to be transported to other storage facilities or further distribution.

As you now understand, in this entire process, every step is critical, as each step brings the raw material closer to the desired end product. Similarly, to generate value with data, we must first understand how data is augmented at every stage of the data management value chain.

Value of data increases with each step of the Data Management Value Chain



The first stage is **data acquisition**, where the organization identifies the right or most useful sources for data collection as per business requirements and establishes the right processes to capture different types of data in the right structure and format. At this stage, the data requirements are properly defined, reducing unnecessary data traffic and eliminating wasteful utilization of storage and computing power. This is like the process of identifying the crude oil wells, setting up the process of extraction and transportation to the refinery.

The next step is **data ingestion**, the process of importing or loading data into a system or data storage infrastructure for further processing and analysis. Data can be imported in various formats from a wide range of sources like APIs (application programming interfaces), streaming data, social media feeds, etc., and can be collected either in batches or micro-batches or streams. This is the stage where data is processed and transformed to bring it into the appropriate format to be stored in desired systems. It is similar to the distillation process where the oil is heated and condensed into various fractions that separate out in the distillation tower.

The next stage is **data enrichment**. It is the process of enriching existing data with additional information or attributes to provide more comprehensive and valuable data sets. This process improves the data quality which adds more depth and completeness to the analysis and insight-generation process. It is similar to the conversion process where, based on the nature of the components, they go through further processes of conversion like cracking, coking or reforming. It typically involves breaking down large hydrocarbon molecules, rearranging molecular structures and removing impurities to improve the quality and properties of the resulting products.

Then comes the stage of **data storage**. At this stage, data, which has been treated for quality and transformed into consumable data sets through amalgamation of various types of data from multiple sources, is sent or uploaded into large-scale data storage spaces like data warehouses, data lakes or data lakehouses. At this stage, the data is curated and ready for use by the various teams for analysis and insight generation. Now imagine large tank farms that provide a centralized location specifically designed for the storage of large quantities of liquids, such as petroleum products, chemicals, liquefied gases and other industrial fluids, to be transported through ships, barges, pipelines, trucks, etc.

The next stage is **insight generation**. At this stage, various statistical modelling and ML techniques are applied to generate insights, and then these insights are made available using business intelligence (BI) tools or self-serve tools, ready to be consumed by the decision-makers. At this stage, data enables deeper and better understanding of the business problems and recommends actions based on these insights. This is much like the blending process, where various components are mixed, 'blended' together, to get the right composition of data called insights. Although the majority of the blending process happens before storage in the refinery process (unlike in our data management cycle where storage happens before insights generation), often tank farms also include facilities for blending and mixing liquids.

The next stage is **orchestration**. This stage is critical to the process of translating insights into actions. It involves the

coordination and automation of various tasks and workflows aimed at executing the necessary actions derived from the insights. The orchestration stage in the context of data insights delivery can be compared to the pipelines established for transporting various oil products. Just as pipelines efficiently transport different types of oil products to their intended destinations, the orchestration layer automates the seamless delivery of insights generated, to internal and external user engagement systems.

The next one is the **action or consumption** stage. At this stage the insights are fed into systems based on which some action takes place that would create the target impact for the business. Just like in an oil refinery, appropriate products are then used as raw material to create other products or power up machines like automobiles, engines, etc.

And finally, the outcome of the data management value chain is the **impact**. It is the tangible result that you achieve through the entire effort. It could be an internal business outcome or a customer impact. We will discuss it in detail a bit later in this chapter. But the impact of data can be equated to the outcome from oil as well. It facilitates transportation, helps power up machines and build new products like plastics, chemicals, etc.

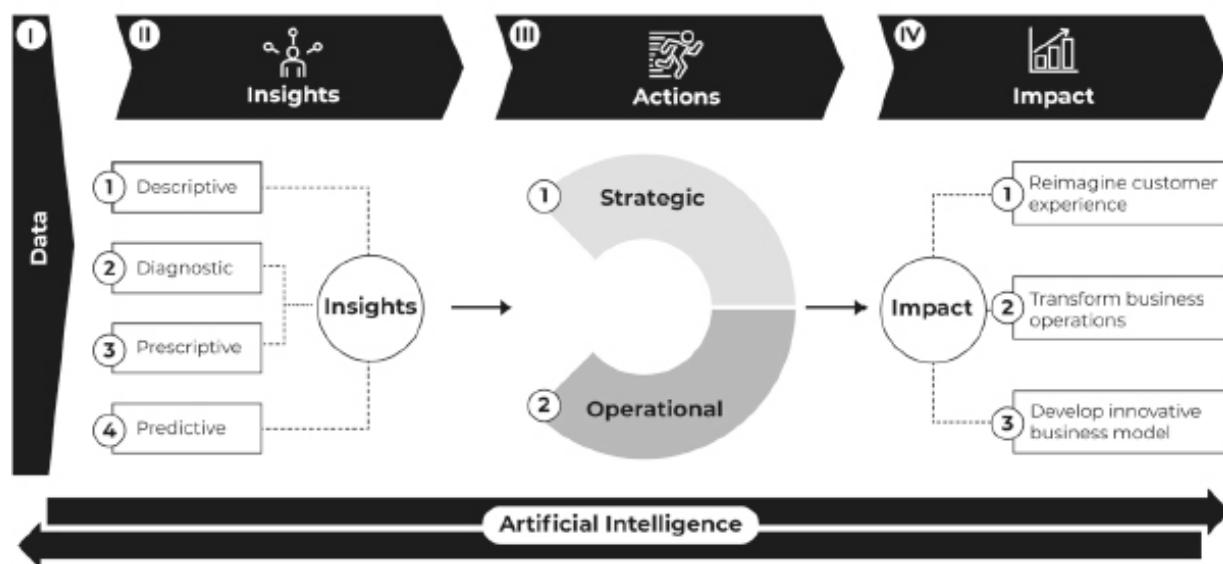
Now that you understand how data is augmented and managed across the data management value chain, let's ask another question. What is the value generated from it? The DIAI framework will help answer this question.

Data-Insights-Actions-Impact (DIAI): The framework for generating value from data

You now understand how data at each stage is processed to enhance it as we move down the value chain. Now across this data management value chain, value is generated at multiple levels. To start with, data itself is a source of tremendous value, which acts as a foundation to drive desired impact. Next comes the value from insights generated, at which you can see some tangible result in

terms of understanding the problem at hand. Action is value generated in terms of decisions taken based on the insights. And finally, with impact, the value is generated in the form of outcome, or the effect of actions taken. Here, AI plays a critical role as an enabler at each stage of the DIAI framework. It helps automate and improve the efficiency, facilitating seamless coordination and enhancing the overall effectiveness of the process.

DIAI framework for maximizing value realization from Big Data



To better understand the value generated at each of these stages, let me take you through the DIAI framework.

I. 'Data' in itself is a source of value

Data serves as a foundation to deliver business value. Data has the potential to uncover patterns, reveal trends and provide deep insights into customer behaviour, market dynamics, operational efficiency and more. By effectively collecting, managing and analysing data, organizations can make informed decisions, drive strategic initiatives and gain a competitive advantage.

- 1. Leveraging multiple data sources:** In the world of Big Data, organizations have access to varied types of

data, which when combined well can enable organizations to uncover deeper and more comprehensive insights. For example, data from multiple sources, which I am going to talk about in detail in Chapter 8, Multi-Source Data, enables organizations to gain a more complete and holistic view of their operations, customers and market landscape, to build a more comprehensive picture. Real-time data is another source that is highly valuable for organizations. This type of data, obtained from sensors, social media feeds or IoT devices, allows organizations to monitor events, trends and customer behaviour in real-time. It enables organizations to shift from reactive to proactive monitoring and response to business requirements by automating decision-making or actions. I have talked about this in detail in Chapter 9, Real-Time Data. Furthermore, another important type of data that adds significant value to organizations is Proprietary Data which is the codified tacit experiential knowledge, developed over time and converted into unique assets that can become a source of differentiation for them in the long run—a topic that I discuss in detail in Chapter 10, Proprietary Data.

2. **Value of data integration:** While it is not an easy task to integrate these various data sources, consolidating and harmonizing data from various sources adds significant value to organizations. By integrating data from different sources and disparate systems, organizations can eliminate data silos and gain a comprehensive understanding of their operations, customers and market dynamics. This enhanced data visibility and accessibility enable more accurate and timely decision-making and actions. Data integration also facilitates advanced analytics, predictive modelling and ML by providing a holistic data set for training and analysis—an essential component for scaling AI efforts.

3. Value of data augmentation: Data augmentation is a way to enhance existing data and make it more suitable and valuable for AI/ML models. It involves various methods to modify and expand the data set by adding missing values, transforming existing data and artificially generating new data. Data augmentation addresses missing value through techniques like imputation—where missing values are filled in based on statistical methods or predictions. It transforms existing data sets by enhancing variability by making it more diverse and representative of different scenarios, patterns and characteristics using various techniques like scaling, normalization, etc. And artificially increasing the amount of data by generating new data samples through techniques such as data synthesis, where new data points are generated based on the existing data distribution, which improves the accuracy and predictability of AI models.

II. 'Insightful' value from data

The next level of value is generated with insights. Examining and analysing the data that the organization creates and collects on a continuous basis, can help them answer questions, identify significant trends, patterns and generate insights critical for business decisions on a day-to-day basis or on a strategic level. Various tools, softwares and frameworks can be used to analyse data from multiple angles and create visualizations that can help make sense out of piles of raw data. Any business owner or decision-maker depends highly on data analytics to make informed decisions about their business.

But insights are not just of one kind. There are four types of insights that are generated through data analytics, whose use depends on the type of decision to be made.

1. **Descriptive insights:** As the name suggests, this is the basic type of analytics that describes a business situation and forms the basis for other types of insights to be built on. This type of analysis answers the question, 'What happened or is happening?' Take the example of an e-commerce company. And I will stay with the same example for this section to bring out the difference in various types of insights. They can analyse the customer's buying habits, most popular products, frequency of purchase, or which products have high return records. Of these, let us narrow it down to the return example. The company wants to understand the pattern and find an effective way of dealing with it. Upon analysis of the returns' data, two issues are highlighted. First, the returns are high during the festive seasons and two, they are high on certain products. Bulk of management reporting is such descriptive insights that present 'what happened' and are often delivered in standard reports.
2. **Diagnostic insights:** Diagnostics is the next level of insight, delving into the 'why' behind data patterns. It involves uncovering correlations, causal relationships and coexisting trends to understand root causes. For instance, for the e-commerce company under discussion, high return rate issues can be analysed further by studying customer complaints, product types and brands. It is diagnosed that clothing products, particularly certain brands, are being returned during festive seasons, due to quality issues. Diagnostic insights provide a focused understanding of specific problems and serve as a basis for informed corrective actions.
3. **Predictive analytics:** Predictive analytics involves forecasting future trends or events based on historical patterns. The question answered here is, 'What might happen in future?' By analysing return patterns, recent sales data and incorporating variables like customer

demographics and purchase history, the e-commerce company can predict potential return rates more accurately. This insight allows the company to make informed decisions regarding brands experiencing higher returns. Predictive analytics often involves building models that transform input variables into predictions, leveraging statistical analysis and pattern recognition. It's a key component of data science, which has grown significantly in recent decades.

4. **Prescriptive insights:** And finally, prescriptive insights help answer the question, 'What should be done next?' It considers all the possible scenarios and outcomes and suggests the best course of action. By analysing all customer complaints and returned-items data, the e-commerce company can identify the areas of improvement. Solutions like enhancing product descriptions, providing size guides, offering free returns, or try-before-buy options can be evaluated to address the returns issue during peak seasons. Similarly, analysing the customer complaints and return reasons for brands with high-return-rates aids in identifying areas of product improvements, or discontinuation or replacement brands with similar product offerings.

Evolution of Insight Generation—from Analytics to AI

Over the years the scope and scale of analytics has grown tremendously, and it has become a very powerful tool for insight generation. It is being constantly leveraged across organizations to optimize performances. And data, with its growing complexity is posing a significant challenge for organizations to draw meaningful insights.

To cater to this evolving need, the world of analytics has witnessed four prominent shifts, which I have also talked about in my previous book, *Winning in the Digital Age*. Analytics is moving towards a prescriptive mode from the descriptive, thanks to machine learning enabling the shift from insight to action. As a result, analytics has become a front-line capability rather than a support. Analytics is also becoming a continuous process, shifting from batch processing that was done offline to online continuous analytics and insight generation, based on real-time data.

Analytics is also augmented by combining AI and ML that enables automation of tasks that were previously done by the data scientists. And finally, cloud-based analytics provides scalable infrastructure for computing capable of supporting Big Data and real-time analytics.

So, the world of insights generation has witnessed a prominent shift from offline, batch processing to online real-time insight generation and the decision-making process has shifted from asynchronous to synchronous. And while the nature of insights remains the same—the four types described earlier—the process of insight generation has moved from humans to machines. Primarily because the data has also transitioned from being generated in batches to real time. And as this shift is happening, we are moving away from the paradigm of analytics to AI. As an example, the Data and Analytics team at Incedo has now been renamed ‘Data and AI’ team. This is because I believe there is a decisive shift happening in insights generation from analytics to AI.

III. 'Actionable' value from data

AI—Automating data to insights

The insights generated in the earlier stage of the DIAI framework should lead to actionable recommendations that require some data into actionable insights. Traditionally, extracting insights from decisions to be made and eventually actions to be taken. A business data involved manual data processing, analysis and interpretation, is run as a result of a range of actions taken by multiple stakeholders. These actions fall on a spectrum that ranges from AI, this process can be revolutionized.

strategic actions—the big-picture decisions that can determine the course of a business—to the operational actions that determine the day-to-day working of the business.

- **Enable scale:** AI-powered tools can analyse vast amounts of data quickly and accurately, providing

valuable insights about customer behaviour and

1. **Strategic actions:** Strategic actions are the big-picture, preferences. This includes analysing data from organizational level actions that have a long-term impact on the organization. These actions are programmes or multiple sources such as the organization's internal data, industry data, social media data, customer reviews, etc. For example, social media platforms projects that help organizations achieve their long-term goals or objectives. Since these decisions are generally like Instagram use AI at scale to generate insights taken for a longer term, the level of ambiguity is high, on user-generated content.

- **Higher accuracy:** By using AI tools to automate the insight-generation process, businesses can reduce the need for human staff to analyse data manually, avoiding human errors. Moreover, AI has strategic decisions on how to achieve this. For example, the ability to unearth hidden patterns in the data which might be difficult for organizations to targets. Data can help understand the organization's current carbon emissions and set achievable reduction targets. Supply-chain optimization can be done by on patients and physicians, and combining these with electronic medical records and demographic information more accurately, AI algorithms can help to identify areas for reducing emissions throughout the sales representatives customize their messaging for supply chain. The organization will also need to specific health care professionals.

- **Insight generation in real time:** Real-time insight generation enables businesses to make decisions and take action in real time. AI tools such as reforestation or investing in clean-energy initiatives.

2. **Operational actions:** Operational decisions or actions leverage AI algorithms to generate insights from are the day-to-day or short-term decisions that help customer transaction data in real time to identify

execute strategic transactions. These decisions typically involve less insight generation and need to be taken in a shorter time frame.

Typically, insight generation from data has been limited to operational analytics teams or specialized AI systems. But AI-powered (especially generative AI) digital platforms and tools can enable organizations to achieve their strategic goals. For example, as part of the goal of becoming carbon neutral, the organization must make several operational decisions such as fitting AI into energy management, where insights from AI can help open up the data of the world for everyone to leverage and generate opportunities specific to their requirements. This could include optimizing energy use, reducing waste and implementing energy-saving measures. Or decisions related to transportation planning to optimize routes, minimize fuel consumption and encourage the use of electric or hybrid vehicles. The organization would also need to track and record carbon emissions, where data analytics can help ensure accurate measurement and reporting of carbon emissions.

Another critical aspect to achieving the goal of carbon-neutrality is continuous monitoring and feedback. Monitoring through real-time data provides ongoing feedback on the carbon-reduction process, critical to stay on track or make required adjustments to the strategic decisions.

AI—Automating insights to actions

While it might be difficult for AI to automate strategic decisions, it is highly effective in automating operational decisions. Once AI systems generate meaningful insights, they can be connected with operational processes and systems to drive actions. AI algorithms can analyse and interpret the insights in real time, enabling immediate responses and automated decision-making. A few ways in which AI can translate insights into actions automatically are:

- **Intelligent automation:** AI systems can automatically analyse and interpret data insights, identify patterns and trends, and determine the most effective actions to take. Through integration with various systems and processes, AI can initiate and execute actions autonomously, eliminating the need for manual intervention. For example, a customer service chatbot that utilizes natural-language processing (NLP) and ML to understand and respond to various customer queries, provide relevant information, troubleshoot common issues and even escalate more complex cases to human agents when necessary.
- **Automate workflows:** AI can automate business workflows by routing tasks, assigning resources and optimizing processes, without human intervention. It can be used to analyse incoming requests, distribute work to the appropriate team members and automate follow-up actions or notifications. For example, in an automated warehouse, AI can optimize the picking and packing process. Through data analysis, it determines the most efficient routes for picking items from shelves and guides warehouse robots to fulfil orders quickly. If any anomalies or discrepancies are detected during the packing

IV. 'Impactful' value from data

Impact is the final stage where the value of data is realized.

Seemingly mundane sets of data have the potential to make a transformational impact for business. AI also plays a critical role in driving transformational impact by automating entire processes and enabling new capabilities for end customers. Data can potentially transform businesses in three major ways:

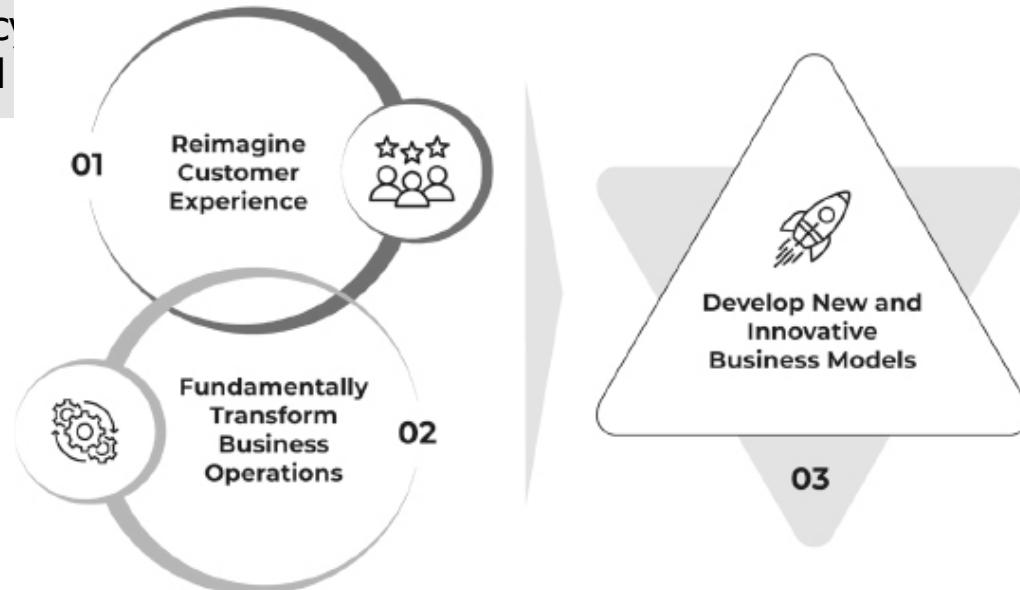
making action without the need for constant communication with a remote server. By deploying

1. Reimagine customer experience

In the world of connectivity, customers, who seek Net neutrality and privacy, and increased companies must security. A good example is their autonomous car that utilizes AI to analyse data from sensors, cameras and maps to navigate safely and efficiently. The AI system processes real-time data to identify obstacles, recognize traffic signals, augmented by AI, have changed companies to build personalized driving and steering all without like never before. I have talked about this in detail in Chapter 9, Real-Time Data.

Data can help organizations reimagine the impact that can be generated

By automating efficiency, respond



Many trends have converged to help deliver high-quality personalized experiences to customers. Organizations can deliver an **omnichannel engagement** by seamlessly integrating multiple channels across online, offline and hybrid. They can hyper-target their customers through **hyper-personalized recommendations** by leveraging granular data available on each customer. Organizations can also provide **frictionless customer experience** by monitoring and analysing the customer journey to eliminate bottlenecks. The B2B players, who are now expected to deliver customer experience at par with any B2C company, are also able to establish newer and more **innovative standards of experiences** for their customers.

The Amazon example is well known; we're all aware of how data reshapes B2C customer experiences. However, a less discussed yet significant trend is the growing impact of data on B2B customer interactions. B2B enterprises are now beginning to be conscious of their client's experience similar to end customers in B2C models. Consider an aircraft manufacturer whose customers are airlines. Ensuring excellent customer experience involves post-sales monitoring and maintenance. Prioritizing predictive maintenance to offer real-time, proactive solutions that minimize costs, downtime and operational disruptions for their clients can be a significant boost for their customer experience. And effective use of data can make this happen.

Take Boeing as an example. It is a pioneer in using Digital Twin technology—a virtual representation of a physical object, system or process, created by collecting real-time data from the physical counterpart and using that data to build a dynamic digital model. This enables them to monitor the performance of various components of the aircrafts on an ongoing basis without the need to

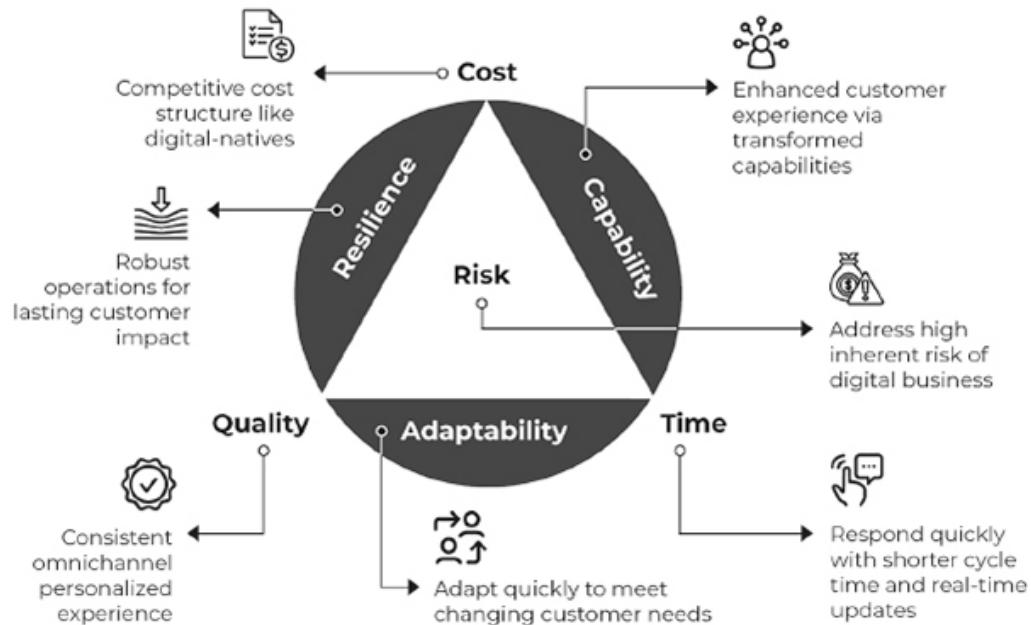
be brought in for physical inspection. Integrating the data collected from sensors placed on physical components or systems with the corresponding digital representation in a virtual model enables the company to predict maintenance requirements proactively and take corrective measures in time. This results in safer operations and reduced maintenance costs, both outcomes are critical for airlines.³

2. Transform business operations

Traditionally, business operations were considered to have created value if they delivered on four dimensions—delivering **quality** products, in reasonable **time**, at lowest possible **cost** while managing **risks**, both internal and external. But today, businesses operating in the VUCA world need to deliver not just on the above four dimensions, but also be more agile in handling disruptions and should be able to leverage multiple capabilities across their value chain simultaneously to deliver value seamlessly. As a result, three additional dimensions continuously come into play—capability, adaptability and resilience. Now companies must build **resilient** operations that are able to resist, adopt, absorb, recover, or adapt to any adverse event or occurrence due to external or internal factors.

Furthermore, the operations must be easily **adaptable** to the changing conditions and business dynamics since the business problems keep evolving very rapidly. And finally, the operational **capabilities** must enable the company to build new competencies to deliver enhanced or unique customer experiences. Big Data enables companies to do so effectively.

Big Data enables businesses to go beyond the traditional dimensions of value creation in operations



Take UPS, an American multinational package delivery and supply chain management company, as an example. By 2019, they were handling around 5.5 billion packages across 11 million customers. Their delivery fleet grew to about 1,25,000 package cars or vans and advanced technology vehicles, and around 500 owned and leased jets. Operational flexibility and efficiency became paramount. On-Road Integrated Optimization and Navigation (ORION), their proprietary route optimization software, solved their biggest challenge around route optimization using fleet telematics to gather real-time data.⁴ Data like vehicle location, engine diagnostics, vehicle activity combined with driver behaviour, geospatial data and weather conditions were analysed through advanced algorithms to deliver tens of thousands route optimizations per minute and suggesting route alteration in real time in case of any adverse weather conditions. Network Planning Tools (NPT) to optimize the flow of packages and Enhanced Dynamic

Global Execution (EDGE) to optimize operations, further streamlined processes using real-time data, AI and analytics.⁵

3. Develop innovative business models

Big Data coupled with new technologies like AI, ML and IoT, has brought about significant disruption and realignment in traditional business models. Those, we call the 'disruptors', have leveraged the abundance of data to build transformative business models that completely redefine customer experiences while often taking an innovative approach to business operations. By disrupting both the 'what' and the 'how', they have succeeded in creating innovative business models. Digital natives like Google, Amazon, Flipkart, Netflix, Uber, Airbnb and many others, though operating in different industry spaces, have all created amazing, new business models that have data at their core. It is not an exaggeration to say that these tech giants are data companies at heart!

Essentially, they are trying to leverage and monetize data in unique ways. It ranges from selling raw data converted into intelligence using data to create differentiated offerings, to creating networks that can deliver data as, when and where needed. These businesses have been able to uncover newer insights, reach untapped customer segments and build innovative new products that have changed the game altogether. Multiple business models have emerged over the years, but these can largely be categorized into five types:

- **Data as a service (DaaS):** DaaS, is a business model that involves curating, aggregating and meshing data from multi-sources to offer value-added intelligence or information to customers. S&P Global is one such

company that provides content and analytics capabilities, primarily specializing in financial information and analytics. Thirty-nine per cent of S&P's total revenue (\$3.25 billion) comes from their data subscription business.⁶

- **Ad-based model:** A model in which digital businesses generate income from advertising, where a fee is paid by the advertiser to the platform provider for access to the audience. As I discussed in the previous chapter, most social media platforms like Facebook, X (formerly Twitter), etc., have adopted an ad-based business model. The Google Ads platform enables advertisers to display ads, product listings and service offerings across Google's extensive ad network of properties, partner sites and apps, to web users. Google Ads earned \$224.47 billion in 2022.⁷
- **Marketplace:** A digital platform that connects buyers and sellers of products and services, where marketplace platforms don't necessarily sell any products or services of their own. Most e-commerce companies follow a marketplace business model where they sell products from other sellers, in addition to their own brands. A good example of a marketplace model is Amazon's third-party seller services that provides a platform to individuals and businesses who want to sell their products on the platform. Amazon's marketplace services yielded \$118 billion in revenue in 2022, constituting 23 per cent of total sales and showcasing double-digit growth from the previous year's 21 per cent.⁸
- **Aggregator:** A single, uniform platform for the user that brings together often unorganized and/or highly diversified service providers, under its own brand and is responsible for the quality of products and service standards. Like Uber, a technological platform that

connects drivers and passengers to provide cheaper and easy transportation to the customers and a source of income to the drivers. It made \$14 billion from ride hailing services in 2022.⁹ Aggregator business model has been one of the most successful innovations over the past decade where Uber is the largest car-ride service in the world without owning a single car and Airbnb is one of the most valued hotel companies in the world without owning a single hotel room!

- **Direct to customer (D2C):** An important shift that I have talked about in Chapter 1, Data Explosion, and in my previous book as well, is disintermediation. A business model where manufacturers/producers sell their products online directly to end-consumers eliminating the middlemen/distributors from supply chains. Almost every industry is leveraging data to better understand their customer to make this shift. Like Tesla which completely revolutionized the idea of selling cars by eliminating the traditional car distributor channel and directly selling it to the customers, just like any other electronic product available online. Its revenue grew by a whopping 51 per cent in 2022 over the previous year, reaching \$81.5 billion and its market value is multiple times more than that of Ford and GM combined though it makes fewer cars. ¹⁰

Generative AI: The next frontier to deliver transformational value

As discussed earlier in this chapter, AI has the potential to make a significant impact across Data-Insights-Actions-Impact (DIAI). However, it is generative AI that holds the transformative power to revolutionize each component of the framework. Let's see how.

Data: Gen AI plays a crucial role in bridging the gap in data availability by generating synthetic data—artificially generated data that mimics real-world data patterns and characteristics, created using algorithms or statistical models to replicate the statistical properties and structure of actual data. This is particularly valuable when real data is scarce, expensive to obtain, or subject to privacy constraints. NVIDIA's Omniverse Replicator enables autonomous vehicle developers to leverage physically accurate or realistic synthetic data, to speed up the training programmes by allowing algorithms to continuously assess their performance in a completely digital environment, 24x7.¹¹

Insights: Gen AI models built on data of the world, possess broader predictive abilities compared to those built on limited organizational data. By building business-specific models on foundational models, organizations can adapt AI to their specific needs and domains. Alternatively, training foundational models on deep proprietary data taps into internal data and expertise, generating contextually relevant insights for a competitive edge. Morningstar, a prominent investment research firm, created Mo, a research tool utilizing Gen AI. Aimed at aiding financial advisors and individual investors, Mo was trained on over 10,000 pieces of Morningstar's exclusive research. In its initial month of operation, Mo effectively addressed 25,000 queries at an average expense of \$0.002 per question.¹²

Actions: Gen AI enables persona-based (enterprise or individual) contextualization of insights, significantly improving the chances of generating actionable insights. Gen AI models can mimic human responses based on text prompts making it extremely easy to use and can drive high levels of automated actions. In collaboration with IBM Consulting, Bouygues Telecom is leveraging enterprise-level Gen AI capabilities to revamp its call centre operations. Human operators were struggling to capture customer interactions comprehensively and read lengthy call transcriptions to take actions in real-time. But

now, automatic call summarization and topic extraction along with accurate and instant actionable insights are being delivered to the agents by leveraging IBM's foundational models. This innovation has led to operational improvements yielding savings of over \$5 million and a notable 30 per cent decrease in pre- and post-call operations.^{[13](#)}

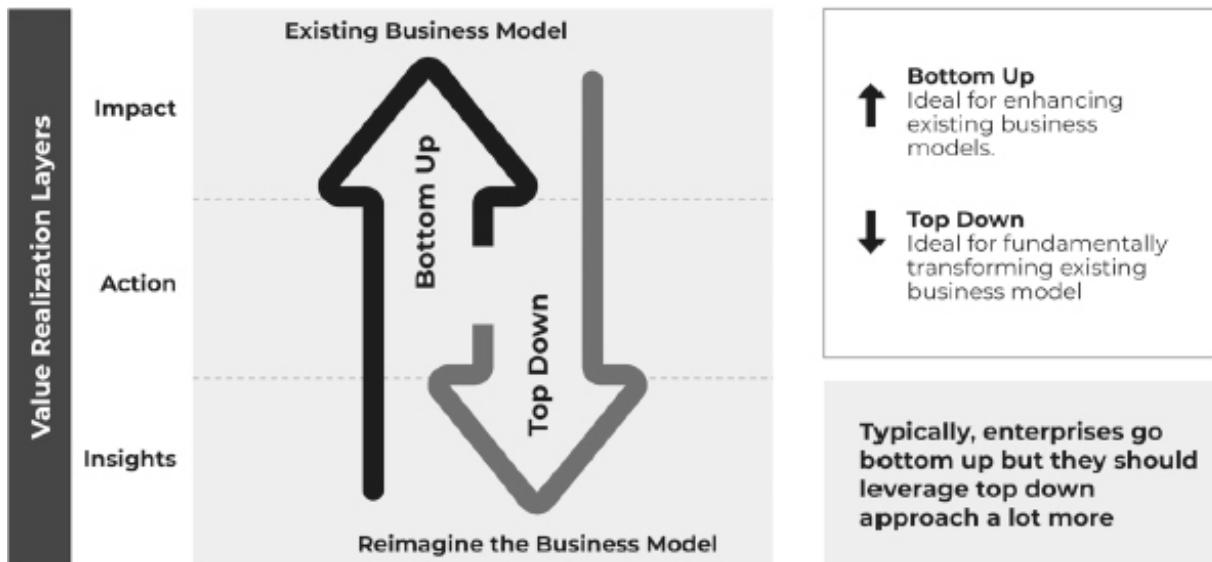
Impact: By leveraging the vast amount of internet data, Gen AI offers organizations access to complex and sophisticated AI models that would otherwise be challenging to develop in-house, except for a few digital natives. These models provide organizations with a strong foundation to build upon, to deliver exceptional customer experiences, optimize business operations and create disruptive business models. A recent study by McKinsey estimates the potential economic value generated by Gen AI to be between \$2.6 trillion and \$4.4 trillion across industries.^{[14](#)} For example, Google applied fine-tune training to enhance its Med-PaLM2 (second version) model, focusing on medical knowledge. They started with the foundational PaLM2 large language model (LLM) and retrained it on carefully curated medical knowledge from a variety of public medical data sets. As a result, its accuracy in answering U.S. medical licensing exam questions reached 85 per cent, almost 20 per cent better than the first version of the system.^{[15](#)}

Putting the DIAI framework into action: Starting a data-led business transformation journey

Now that we have understood the DIAI framework for generating value from data, the question is how do you put it into action and get started with a data-led business transformation journey? The beauty of the DIAI framework is that it can be leveraged both ways, bottom up or top down, depending on your situation and the nature of impact you want to create. Value can be realized both bottom up, starting from building a foundation with insight generation, and top

down, starting with the impact you want to create and working backwards.

Data led business transformation can be both bottom up or top down



The bottom-up approach is ideal when an organization is looking to improve its existing business model, make incremental changes to achieve enhanced value from the existing business model.

Walmart has been one of the most successful retail chains in the world. They have used data to drive revenue growth and improve efficiency by taking a bottom-up approach to drive value from data. In 2011, Walmart set up a Walmart Labs team responsible for building their Big Data capability. They focused on building the foundational capability by setting up 250 nodes Hadoop cluster for enabling use of vast amount of data from their customers, and their operations for analytics. Also, they started the process of setting up the largest private cloud with the ability to process 2.5 PB of data every hour. They continued to add newer capabilities and toolsets to model, manipulate, visualize and deliver insights in real time for decision-makers. This effort culminated into the 'Walmart Data Café'—a collaborative analytics facility, which allows all Walmart's employees to access and analyse data quickly and easily, enabling them to make data-driven decisions, any time, anywhere.

This data café has been instrumental in transforming their inventory tracking system, set up in early 2000s, into a world class inventory management solution over the years. By integrating it with data café, Walmart could now analyse sales data, weather patterns and multiple other relevant factors, to accurately predict product demand and optimize inventory levels. This helped them to effectively streamline their supply chain, minimize stockouts and improve overall operational efficiency.¹⁶

Traditionally, organizations have embraced a bottom-up approach, which is what I also see with most of our clients at Incedo.

But the more revolutionary approach, which has been used much more in recent years, and is the need of the hour, is the top-down approach, where you start with a vision for a business model you want to create anchored on data. You start with a business requirement, identify the actions that are needed to be taken and then go down to the insights required for the same. And then you identify the right data required to solve the business problem. For example, Uber identified the need to revolutionize the transportation industry by providing a convenient and efficient way to connect riders with drivers, creating a seamless experience for both parties. Based on this requirement, Uber focused on creating a platform that would serve as a digital bridge between riders and drivers. This involved developing a mobile app that allowed users to request rides and drivers to accept those requests. To fulfil this vision, Uber needed insights into transportation habits, urban mobility challenges and user preferences. It sought to understand how people used taxis, what frustrations they encountered and how technology could improve the experience. Uber's top-down approach led to the creation of the revolutionary ride-hailing platform. By addressing the core business requirement of connecting riders with drivers, Uber transformed the way people access transportation.

If you look closely, you would notice that the legacy organizations have generally taken the bottom-up approach while the digital natives, the disruptors, have gone for the more revolutionary approach, the top down one. And as is evident from the impact they

have been able to create, the digital natives are the ones who have been more successful in realizing transformational value from data. So, in a world where data is exploding like it is, and business dynamics are changing so rapidly, the bottom-up approach is not enough. It is time legacy organizations rethink their approach to generating value from data or they would be at risk of being left significantly behind in the data-first world.

Key takeaways

- Although most organizations recognize the potential of data to create value for their businesses, they lack a structured approach to maximize value from it.
- Understanding the DIAI (Data-Insights-Actions-Impact) framework can be a helpful guide to maximize value realization from data. At each stage of the DIAI framework, the value realized is augmented manifold, with maximum value realized at the impact stage.
- There are tremendous opportunities to create new, innovative business models that have data at their core. Success stories include digital natives like Google, Amazon, Flipkart, Netflix, Uber, Airbnb and many others, which are data companies at heart!
- AI is transforming and creating new opportunities across the various stages of the DIAI framework, speeding up and enhancing the process of transforming raw data into desired impact. Generative AI can further revolutionize each stage of the DIAI framework.
- An organization can choose to embark on its data-led business transformation journey going either bottom up or top down on the DIAI framework. Digital natives have often followed a top-down approach and driven breakthrough innovation. Traditional organizations need to reconsider their incremental, bottom-up approach to remain competitive in the data-first world and AI age.

The Data Paradox

Deluge and Drought

'How wonderful that we have met with a paradox. Now we have some hope of making progress.'

—Niels Bohr,
Nobel Prize winning physicist

Data has humongous potential to generate value and has been generating unprecedented value for years for us. I hope Chapter 2, Data, the Fuel for the Digital Age, has made that abundantly clear. I also highlighted that organizations need a systematic approach to drive value from data. In fact, one survey claims that 66 per cent of businesses see themselves as data-driven organizations, but in reality, only 21 per cent treat data as capital.¹

Doesn't this sound very strange? Despite having so much data at their disposal, organizations are unable to leverage it efficiently. Why? Because the 3V nature of data is creating a situation for organizations where, on the one hand, organizations are struggling to deal with the sheer volume of data available to them today, and on the other, they are not able to do much with the data that they already have. This is what I call the Data Paradox.

And that in essence defines the contradictory nature of problems organizations face today, in realizing the full potential of data. So let me dive deeper into explaining this paradox and why I think solving for this challenge is the key to winning in the data-first world.

The paradox created by Big Data

Paradox is a statement, proposition or situation that seems illogical, absurd or self-contradictory, but which, upon further scrutiny, may be logical or true. One of the most famous paradoxical statements is, 'This sentence is false.' If this sentence is true, then it is by explanation false, thus making it true. It's like an endless loop someone is stuck in. One of my all-time favourite paradoxes is the famous lyrics from the iconic song 'Hotel California', 'You can check out any time you like ... But you can never leave.'

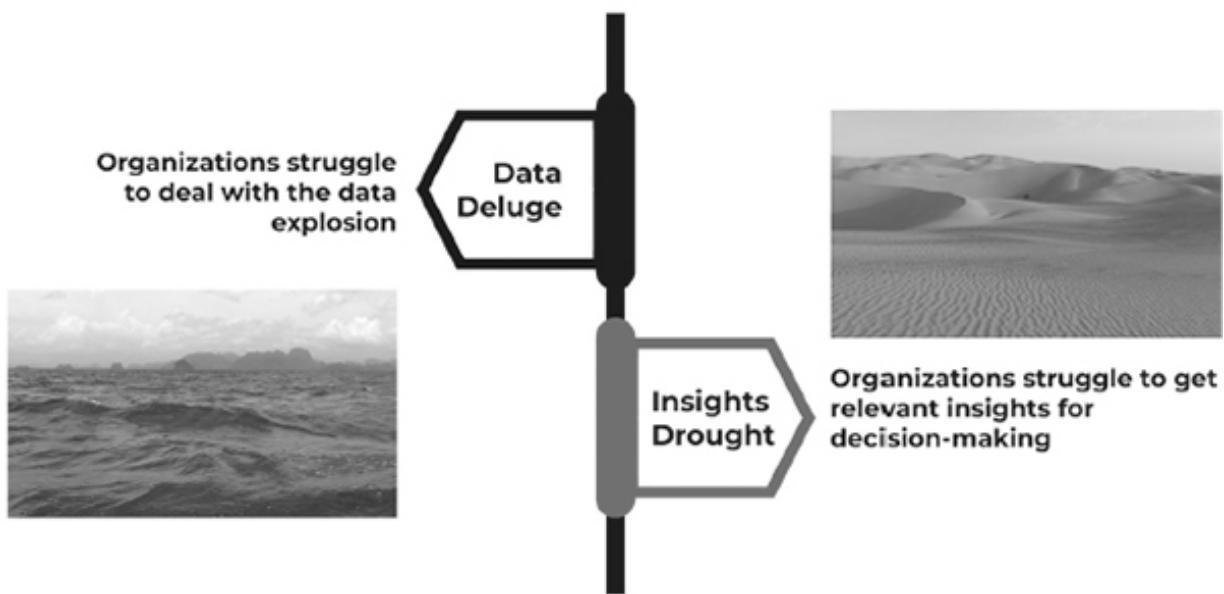
With data that is exactly the case. Big Data is believed to be the answer to all or most of our questions. It is considered a source of meaningful insights that can help organizations steer through the VUCA world that they operate in. But that does not seem to be the case. In a 2021 study, not even half (47 per cent) of the data strategy decision-makers believed that they have been able to enhance the quality of actionable insights as a result of using Big Data.²

Contrary to all expectations, as the data gets bigger and more complex, more and more organizations are now finding themselves in a situation where on the one hand they are awash with data, so much so that they are unable to deal with its 3V nature, but at the same time, they are unable to leverage the data they already have, to drive their decision-making. This is the deluge-drought conflict called the Data Paradox. Data deluge is a situation where the sheer volume of data being generated is not only putting huge pressure on the technological capacities of organizations but is also becoming overwhelming for organizations to deal with while leveraging it for decision-making. It is surpassing the organization's capacity to capture, store and analyse effectively, in a timely manner. With about 2.5 quintillion bytes of data generated every day,³ you can imagine the magnitude of data that is piling up with the organizations. On the other hand, insights drought is where despite having access to so much data, executives are not effective with it. This is because they lack the required level of skills and knowledge

to understand and use the data, and the lack of data-driven culture which results in gut-based decision-making, instead of relying on data and analytics. So, while organizations are dealing with a massive influx of data, executives are unable to draw insights from data they already have, in turn they keep asking for more data; 'Welcome to the Hotel California!' I must say.

A survey reveals that while 67 per cent organizations say they constantly need more data than their current capabilities provide, 70 per cent say they are gathering data faster than they can analyse and use it.⁴ Let me explain how data is creating such an ironic situation for organizations.

Most organizations struggle with the 'Data Paradox'



Data Deluge: Drowning in data

A study suggests that the volume of data is doubling every eighteen months.⁵ 77 per cent of the tech-industry leaders surveyed admit that dealing with the ever-increasing volumes of data is amongst the top three challenges for them and 61 per cent executives, in another survey, say their teams are already overwhelmed by the data they have.⁶

Furthermore, owing to the prevalence of unstructured data, traditional approaches to data processing aren't appropriate. Systems essentially designed to process static and structured data struggle with real-time and unstructured data. Variability of data, for both structured and unstructured data, has also gone up exponentially, owing to diverse internal and external sources, in varied formats. Legacy systems fall short in managing this scale and variety, making data harmonization challenging. And this challenge will continue to become bigger because data has grown at a much faster rate than the technology to deal with it. As I mentioned in Chapter 1, Data Explosion, data has grown by 1,00,000–1,50,000 times over the past twenty-four years, which can only happen geometrically. But if you look at technology growth, the Big Data market has grown arithmetically, from \$7.6 billion in 2011 to \$77 billion in 2023, registering tenfold growth.⁷

Maintaining data quality amid the data deluge is a significant challenge due to the 3V nature of data. As the number of data points, data types and sources multiply rapidly, ensuring data quality becomes a greater challenge. The adage 'garbage in, garbage out' is even truer in the Big Data era. And storing vast amounts of data doesn't guarantee that every datapoint stored will be useful or be used at the right time. The traditional approach to measuring and ensuring data quality falls short of handling the data deluge efficiently. Additionally, with data coming from various sources and more and more people continuously generating and accessing it, ensuring security and compliance becomes a mammoth task. In fact, a 2023 State of Trust survey reveals that 67 per cent respondents believe their business needs better security and compliance measures. Identity and access management (IAM) and data processing have been identified as the two biggest blind spots for their organizations.⁸ This is why I have dedicated an entire chapter, Chapter 12, emphasizing the importance of data quality in building a truly data-driven organization.

So, forget generating meaningful, comprehensive insights from data, organizations are hit by a metaphorical tsunami that can

overwhelm them. In addition to that, it is also becoming harder to identify the right data needed from the vast ocean of data to be able to generate the required insights.

Insights drought: Parched for insight

Contrary to the challenge of dealing with too much data, the other side of the story is the inefficient but more importantly, ineffective use of data that is available with the organizations today. It is estimated that in 2022 alone 97 ZB of data was created, captured, copied and consumed globally.⁹ And it has been proven that organizations that leverage data for decision-making enjoy 30 per cent more growth compared to others.¹⁰ But with all this data available to organizations, businesses are not effective in maximizing the value from it.

Again, as I highlighted in Chapter 3, Value Reimagined, while insights are generated in abundance, not many are actionable. This is because insights are often generated without involving the business owners in clearly defining the business problem, and therefore insights generated often do not provide the right solution. Let me draw a parallel for this with a personal example. Let's say I am invited to someone's house and guided to an impressive bar with an exquisite collection of globally sourced, expensive wines. I politely decline and opt for water instead. Why? Because I am a teetotaller. And that is where the disconnect is. For example, at one of our key clients (a Fortune 100 company) more than 500 analytical models had been developed, but only a handful were put into production. In fact, a survey reveals that as many as 87 per cent of data science projects never make it into production.¹¹ Which means just one out of every eight projects goes into production.

Additionally, organizations lack a structured approach to understanding how data generates value. While they have so much data at their disposal, being analysed day in and day out, not all of the insights drawn are necessarily actionable. And of the few that are, not all generate the desired impact. As a result, insight

generation becomes a time-consuming and expensive activity resulting in limited impact. In fact, a 2019 estimate predicted that 80 per cent of analytics insights would fail to deliver business outcomes through 2022.¹²

In the dynamic business environment, evolving business challenges demand agility and cross-team collaboration. However, the current data and technology capabilities often lack the required flexibility. Tedious processes around data access and utilization, coupled with siloed systems, hinder collaboration between business, IT and operations. Around three out of five organizations grapple with data silos.¹³ As a result, teams do not have a comprehensive view of the data they need to generate insights they need for timely decision-making. And the incremental enhancements to data infrastructure fall short in creating the required capabilities to draw value from data. A survey found 67 per cent of senior managers are uncomfortable with accessing data via analytics tools, underscoring the need for data literacy and user-friendly interfaces.¹⁴

To add to this, the executives or decision-makers expected to leverage data for decision-making are not adequately trained or empowered and lack the culture that encourages or facilitates effective use of data. Firstly, most business owners with non-technical backgrounds have limited understanding of data's potential and how to work with it. A survey indicates that 61 per cent of organizations lack in-house data science skills and 57 per cent lack in-house technical skills.¹⁵ Secondly, fragmented and siloed data infrastructure makes it challenging for the teams to locate and access the data they need. After all it is estimated that on average, an organization with more than 1000 employees, uses 177 SaaS applications, alongside legacy systems that already exist.¹⁶

Such is the paradox created by the Big Data world!

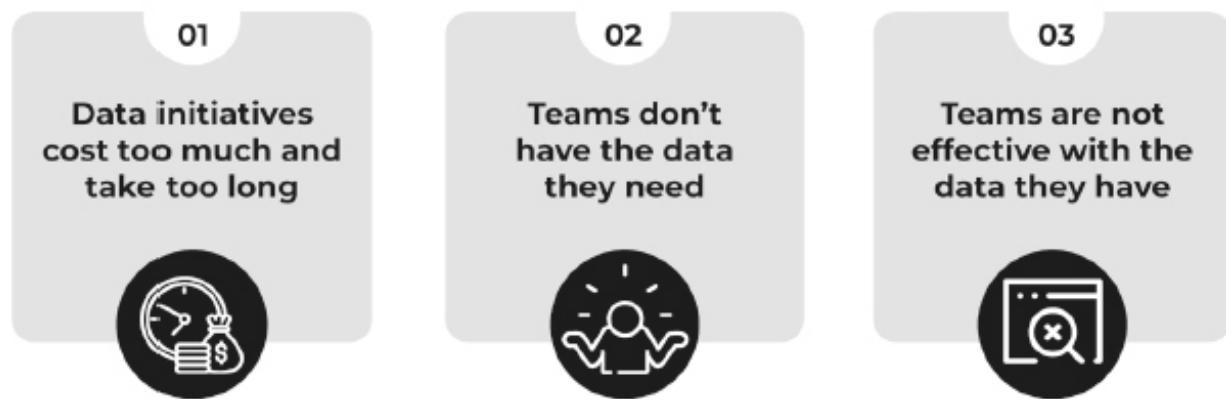
The question is how can an organization recognize that they are stuck in a Data Paradox? During my experience of working with multiple organizations on their data related initiatives, I have

identified certain symptoms that are sure-fire signs of an organization experiencing the Data Paradox.

The three symptoms of Data Paradox

Organizations may display **one of the three symptoms or a combination of any or all of the symptoms** of struggling with a Data Paradox because of which they are unable to generate full value from data.

Three key symptoms of organizations struggling with the Data Paradox



1. Data initiatives cost too much and take too long

Most organizations are working on their digital transformation agenda and as part of that process, are updating or redesigning their data infrastructure to create capabilities that would help them either improve their business process or completely reimagine their business model. Global investment on data and analytics solutions is over \$215 billion annually.¹⁷ For most Fortune 500 companies that we work with, their spend on data is at least tens of millions and many times it is hundreds of millions too. While all this money is being invested in building the right infrastructure to capture, store and process data in the most effective manner, it has been

seen that more than 85 per cent Big Data projects fail.¹⁸ Additionally, it is estimated that close to 60 per cent of all data and AI projects are severely delayed.¹⁹

2. Teams don't have the data they need

A lot of efforts go into capturing, storing and processing as much data as possible, in the hope of making it available to the decision-makers or business owners to enable them in taking the right decisions at the right time and at speed. Despite that most executives feel they rarely ever have adequate data to facilitate data-driven decisions. This is because of all the insights that are being generated, they are often not the ones that the decision owners need. And also, most times, the insights are more of the descriptive nature where the analysis reveals what has happened but lacks insight into why it happened or what may or may not happen as a result. In short, they are not predictive in nature. In fact, only 30 per cent companies are actively using predictive analytics tools while only 3 per cent are using prescriptive

analytics.²⁰ In addition to that, as most of the data is still trapped in silos across different functions or teams in an organization, executives do not have seamless access to it, and therefore they are forced to work with what they have, which is often not enough. While 71 per cent executives feel their data requirements are not always met,²¹ another survey claims that 63 per cent of employees are unable to gather insights within the requisite timeframe, rendering the entire exercise futile.²² Additionally, more than 87 per cent of organizations have been classified as having low business intelligence (BI) and analytics maturity which affects both the teams' efficiency and effectiveness with data.²³

3. Teams are not doing much with the data they have

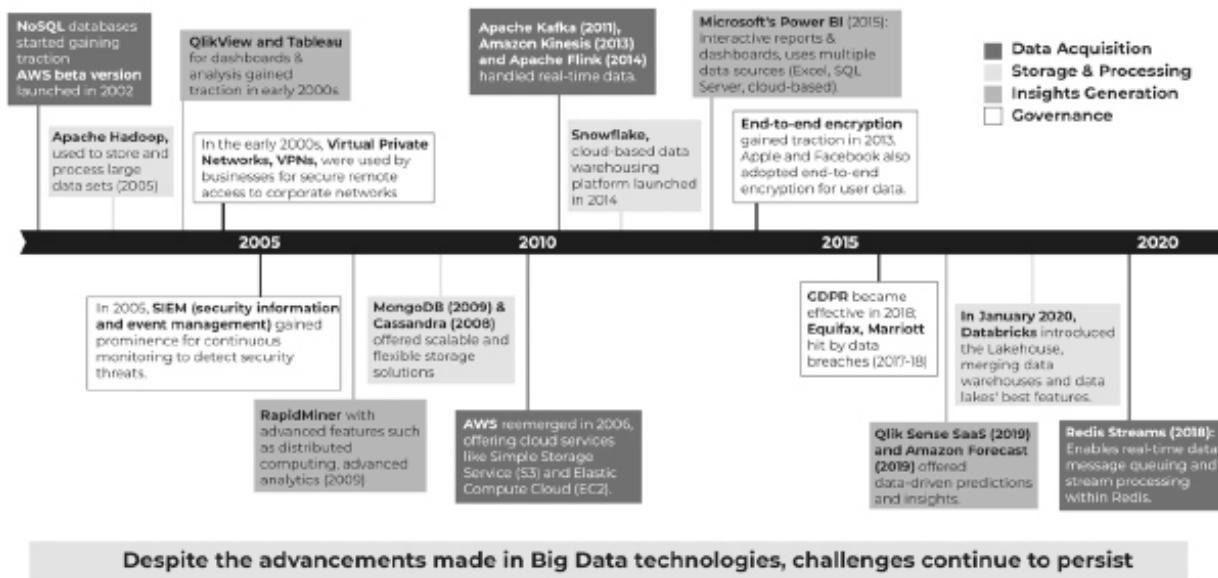
And lastly, with all the data that is available and continuous efforts being made in making the data available to the teams or business owners, they are unable to leverage it effectively to drive decisions and make impact. They do not trust themselves to use the data made available to them. The primary reason for that is the fact that while these insights are mostly generated by a specialized team like Analytics, the business and operations are often not aligned while designing the data approach. This is why the business owners do not feel confident in the insights generated and at times do not understand them either. Another issue is that because the business owners mostly come from a non-technical background, they lack the skills and know-how to understand and use it effectively. In fact, as mentioned before, 67 per cent of senior managers and above are not comfortable accessing or using data from their tools and resources. Lack of seamless and easy access to data and lack of data-driven decision-making culture further leads to sub-optimal use of data. Another significant issue is the difficulty in finding the right specialized data talent that is able to adapt to the changing nature of data and evolving requirements from their roles.

Additionally, the majority of data gathered today, in the Big Data world, is unstructured data and generated in real time so either most of it goes unused or comes with a very limited shelf life, rendering it useless before it can be acted upon. More than half of an organization's overall data is referred to as dark data—data that is collected but never used.²⁴ And equally significant is the fact that even less than 1 per cent of all the unstructured data available to the organizations is being utilized so far.²⁵

Technology alone has not been enough

Obviously, I am by no means the only one who has recognized these challenges with data. It's a situation that many companies have found themselves in over the past few years. And through these years, multiple attempts have been made to find a solution. Many technology companies, including big organizations and start-ups alike, have been developing solutions to tackle the Data Paradox. Let me talk about some of the major advancements over the years.

Over the years, numerous efforts have been made to solve for the Data Paradox



In the early 2000s, in response to the explosive adoption of the World Wide Web, NoSQL was designed to **acquire, process and store large volumes** of unstructured and semi-structured data. Unlike traditional databases, these use **flexible data models** that can adapt to changes in data structures and can be scaled to handle growing volumes of data. But it was only able to make incremental improvements in data management issues.

In 2005, Apache Hadoop* was launched which allowed for **distributed processing of large data sets** in clusters on multiple computers. But managing and optimizing performance for large data

sets can be challenging. During the same period, **BI tools** were introduced but were somewhat clumsy and difficult to use.

Then, 2005 marked the new era of Big Data. Platforms like MongoDB^{*} offered **flexible storage solutions for unstructured data**. But it didn't work well with small data sets and had poor data security. **Parallel computing** gained prominence, enhancing the computational power and enabling advanced analytics. Around the same time, data security started gaining prominence with the advent of continuous monitoring as part of SIEM (security, information and event management).

Real-time data started gaining prominence. **Real-time data acquisition and ingestion** tools like Apache Kafka[†] were developed to handle real-time data. But these tools came with complexities around setting up, configuring and managing them.

Snowflake[‡] was publicly launched in 2014, to provide **cloud-based data warehousing** capabilities. It supports a multi-cloud environment and can integrate well with AWS, Microsoft Azure (Azure) and Google Cloud Platform (GCP). Unfortunately, it doesn't work well with continuous data upload, and unstructured data management support was only launched in 2022.

Databricks,^{*} launched in 2020, which is based on **Lakehouse** concept that provides relational capabilities on top of data lake architecture and provides a platform for data warehousing, business intelligence and data science. While it aims to provide an all-in-one solution, it comes with inherent issues and complexities of integrating multiple data pipelines and managing multiple dependencies. In the analytics space, AI became the defining force. But still scaling those AI solutions has been challenging owing to unavailability of sufficient data sets.

Organizations have been developing innovative solutions to deal with this Data Paradox—from building advanced solutions to better deal with real time and unstructured data to creating a '**data fabric**' that enables end-to-end integration of various data sources with advanced analytical capabilities. While all these technology

developments have pushed the envelope and enhanced the data capabilities, none have succeeded in solving the data challenges fully. The three symptoms that I outlined are still visible in most organizations, the Data Paradox persists and in fact is becoming more prominent by the day. Even after years of technological advancements, a 2021 survey highlights that a massive 83 per cent executives still feel that they are struggling with capturing, managing, analysing or driving actions from data.²⁶

Clearly, addressing three symptoms through technology enhancements, which has been the default strategy for most organizations, has not been enough. I believe solving the data challenges effectively requires us to go beyond these symptoms and get to the core issues. Only by addressing these core issues will an organization resolve the Data Paradox from its root, else these problems will keep resurfacing and organizations will find themselves caught between the rock and the hard place.

We need to dig a little deeper to unearth the real reasons—**the root causes**. So, it's time to turn the page or swipe left, on your devices to find the answer to that.

Key takeaways

- Organizations often find themselves stuck in a Data Paradox: the deluge–drought conflict, which creates challenges in realizing the full potential of data.
- The surge of data across the 3V dimensions is leading to a ‘data deluge’, straining conventional data infrastructure and overwhelming organizations aiming to utilize data for decision-making.
- At the same time, organizations experience an ‘insights drought’ when they possess abundant data but struggle to extract valuable, actionable insights. This can stem from inadequate tools, technology, skills, or a lack of a data-driven culture that promotes data-based decision-making.

- Organizations trapped in a Data Paradox exhibit three notable symptoms. One, their data initiatives cost too much and take too long. Two, their teams do not have the right data they need for data driven decision-making. And three, they are ineffective in leveraging the data they already have, to create desired impact.
- While many organizations recognize these symptoms and have attempted to solve them through technology advancement and investment, the data challenges continue to persist. Clearly, technology solutions alone are not sufficient, and we need to go deeper into these symptoms to understand the root causes.

The Root Cause

More Logical than Physical

'For every effect there is a root cause. Find and address the root cause rather than try to fix the effect, as there is no end to the latter.'

—Anonymous

When consulting a doctor, they carefully listen to your issues, ask probing questions and may request tests to confirm his hypotheses. And only after such a thorough diagnosis do they prescribe treatment. However, we are all guilty of self-medicating at times, because of which symptoms often tend to reappear. Especially with the different kinds of flu—viral, bacteria, allergic—floating around. The symptoms may be similar, but the treatment differs. So why the doctor? Doctors excel at asking the right questions, synthesizing symptoms and identifying underlying causes that we can't pinpoint ourselves.

It is therefore essential to get to the very root of the problem rather than working to solve it superficially. In this chapter, I am going to dig deeper into the three symptoms of Data Paradox to identify the real underlying issues to be able to find an effective solution.

The root cause of the effect

The most fundamental law of science is the 'law of cause and effect'. Every effect has a cause. Or simply put, everything happens for a reason. From this law evolves the law of root causes. All problems arise from their root causes. And if we have evidence that despite

several attempts the problem persists that means we have failed to address the root cause. And the only way is to follow the causal chain to reach the bottom of the problem and identify the root cause.

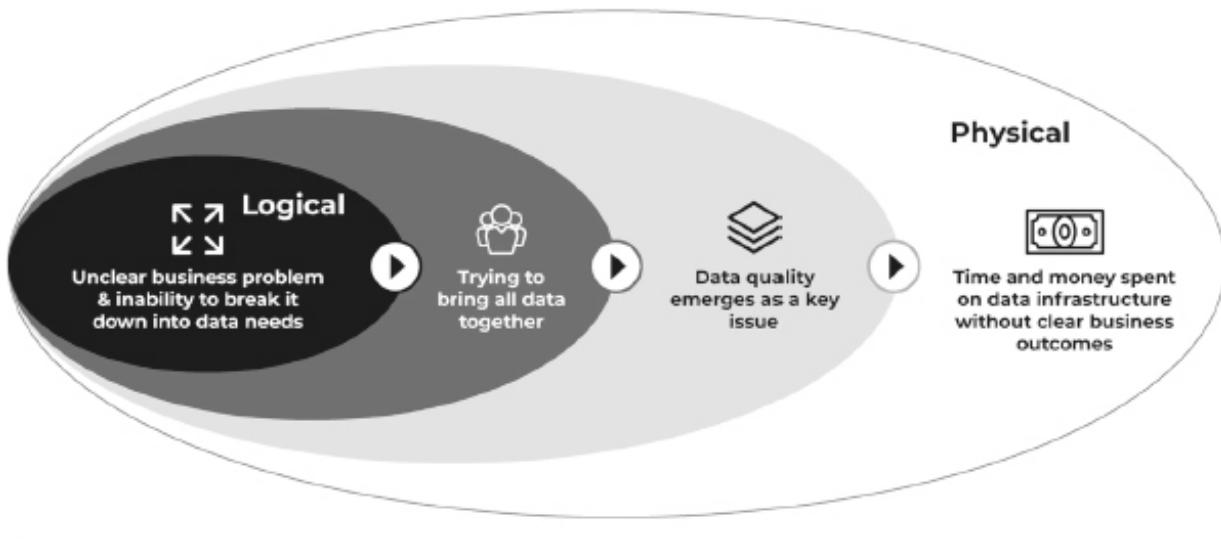
Root cause analysis (RCA) is a tried-and-tested method used for centuries in some form or the other to solve any business problem, rather any problem. The earliest appearance of RCA in a more structured way, was in the field of engineering when the method was first used on the shopfloors of Toyota Industries Co., by its founder Sakichi Toyoda. The '5 Whys' method—ask why five times, was first used to design the manufacturing processes in 1958 at the Toyota Production System. Later on, this method evolved to newer methods like six sigma, KPI trees, etc., but the essence remains the same. ***All problems arise from their root causes and while the manifestation on the surface are so many, the root causes are often very few or one, which once narrowed down become easier to manage and solve the problem effectively.*** So, keep digging until you reach a cause that is fixable and can eliminate the problem from its origin.

The disproportionate focus: Physical vs logical

In my experience, when organizations find themselves facing the data deluge, they often try to resolve it by trying to scale up the data infrastructure. Data infrastructure is expected to be the solution for any and every business problem that the company is facing or may face in the future. And so, they often continue to build massive data infrastructure without clarity on the business problem, and therefore are unable to narrow down their data requirements. As they lack clarity around data required, they start asking for too much data, acting on a common misconception that, if they can get their hands on all the data, bring it all in one place and process it, they would find the solution they are looking for. Unfortunately, when you bring together more and more data, which is of varied nature, and the volume and velocity keeps growing, it becomes very difficult to maintain the integrity and quality of that data. Data quality issues

start to emerge. And to handle this they make huge investments in technology infrastructure that come with the promise of tackling these issues from the word go. And this vicious cycle goes on and on often until it is unmanageable, and you reach a gridlock.

Root cause of Data Paradox: Organizations have disproportionate focus on solving physical issues



Unless organizations focus on the logical issues, investments in the physical issues will continue to result in sub-optimal impact

This disproportionate focus on approaching data with investments in the **physical**—like investing in more storage and computing capacity, without appropriate focus on solving the **logical issues**—by narrowing down the business problem to zero in on the data requirements through collaboration of different functions, results in sub-optimal impact of the data initiatives.

Let us now follow the causal chain of each of the three symptoms that we identified in the previous chapter to distil to the root causes more specifically.

Symptom 1: Data initiatives cost too much and take too long

The fact that I have been focusing on since the start of the book is that in the Big Data world, the amount of data has grown exponentially, expanding the data universe at an unprecedented rate. Organizations today have access to tons and tons of data that is pouring in at an alarming rate. To get the best out of all this data available, organizations are trying to leverage it all to drive their decision-making process, resulting in large and complex data initiatives. But implementing these data initiatives is not as easy as it sounds.

Firstly, organizations are struggling to **integrate and modernize their legacy technology infrastructure** to be able to leverage all the data and bring it all together. But while this legacy technology stack is not designed to handle the scale of data and is inefficient in dealing with unstructured data as well, it is also difficult to integrate with new systems and processes. This issue is further complicated by **existing data silos**, which create a bottleneck both physical and logical to connect all the data. A survey indicates that around 15 per cent of an organization's IT budgets go into maintaining legacy systems,¹ while another suggests that organizations spent an average of 40 per cent of their total IT budgets on IT modernization initiatives.²

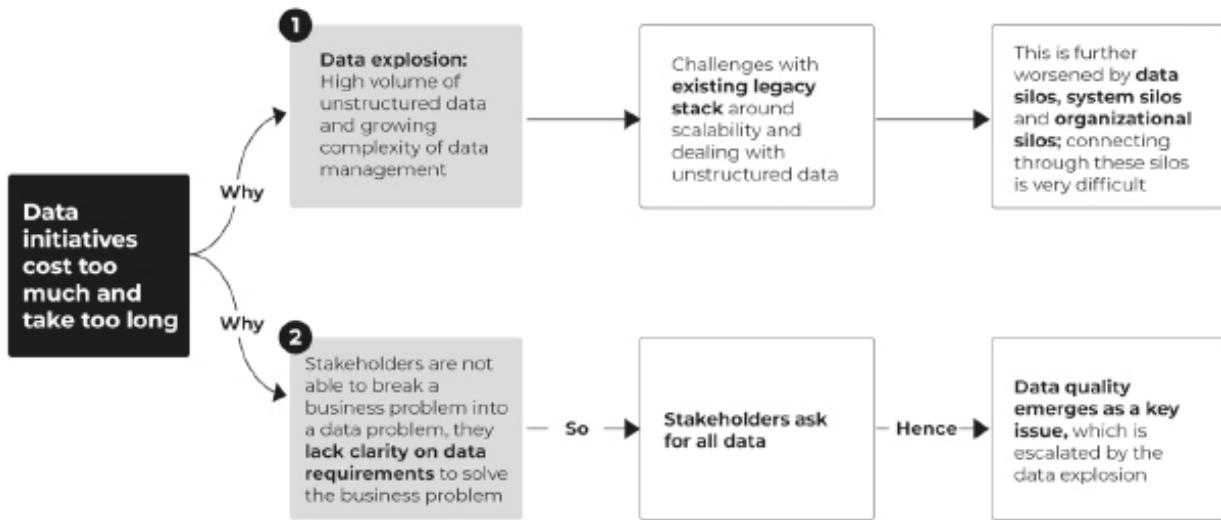
Secondly, the stakeholders or decision-makers are either not putting in the required efforts or are **unable to define the data requirements** properly, or conversations between the data/tech teams and business users on the data that is really required is not happening. As a result, they try to look at as much data as possible assuming data will provide the answer to everything and so end up asking for and **working with too much data**.

Now, when organizations take this bottom-up approach, starting with all the available data, **monitoring and maintaining data quality** becomes a major issue. Why? Because, the very nature of

Big Data, that I talked about in Chapter 1, Data Explosion, the 3Vs, makes maintaining data quality a tremendously difficult task.

Now poor data quality can cost an organization both time and money. It has been estimated to cost companies \$12.9 million annually.³ Apart from direct impact on revenue, poor data quality can keep compounding in the long term to result in poor decision-making or erroneous outcomes. Issues can range from anything as minor as wrong pin code of a customer leading to goods sent to a wrong address, to compliance risks leading to huge data breaches impacting the organizations' reputation.

Issue 1: Data initiatives cost too much and take too long



Symptom 2: Teams don't have the data they need

To start with, the first issue is the lack of relevant insights made available to the teams and decision-makers. Why? Firstly, the insights made available are often not the ones that the business owner needs. The people who are responsible for data processing and analysis don't always understand what the business owner is looking for. This is majorly because of lack of collaboration between the business and data teams. Since business owners are not involved from the beginning in the process of insight generation, the process of narrowing the business problem to define it is not

effective. As a result, most of the times the insights generated are either way off the mark or are too many for business owners to process and identify the right ones. So despite so many insights generated, most of them are not actionable or fail to achieve the desired outcome. Only one out of five data analytics insights were expected to achieve the business outcomes in 2022.⁴ Furthermore, lack of accurate specialized data talent with skills that go beyond core data skills to include storytelling, problem-solving and deeper business domain expertise, which are becoming increasingly important to achieve desired impact. And since this combination is not abundant, organizations struggle to attract and retain such talent. As a result, they are unable to deliver business relevant insights that are easy to understand and leverage by non-technical business users to generate the desired impact.

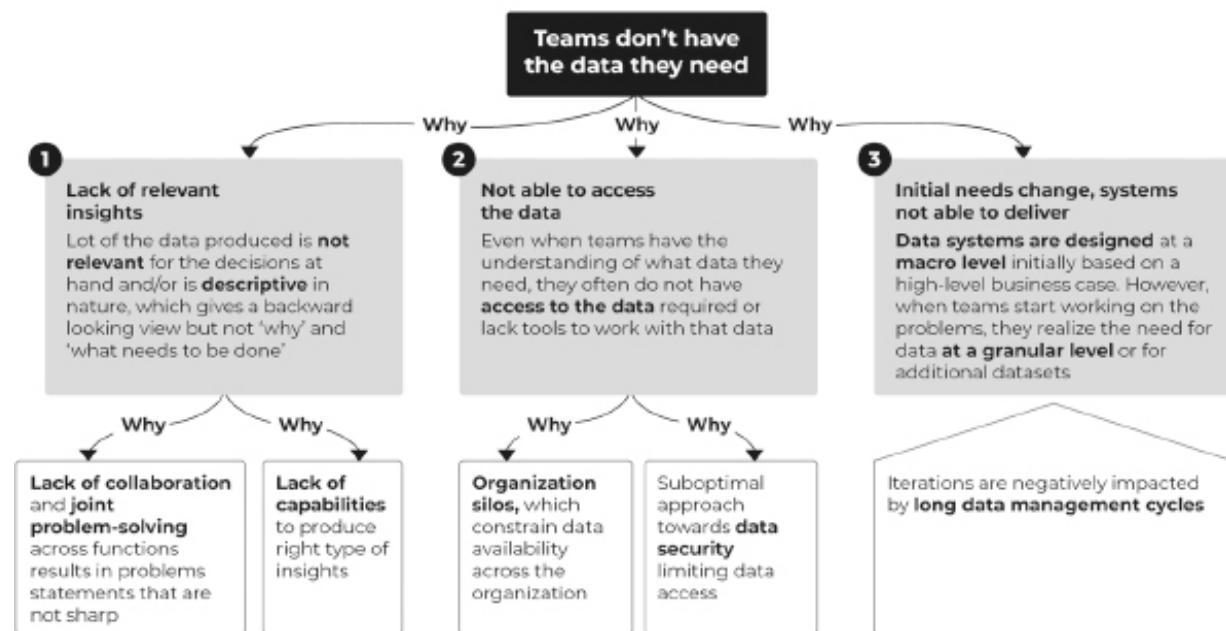
Furthermore, most analytical insights generated are still of descriptive nature, telling us what has happened, but fail to answer why something has happened and more importantly, what should be done about it. And despite high adoption of analytical tools by organizations, most of them are very low on the maturity level. More than 87 per cent of organizations were classified to have low business intelligence (BI) and analytics maturity, in 2018.⁵ Which means most of them are still working with basic levels of insight generation tools. As I also mentioned in the previous chapters, just three out of every ten organizations use predictive analytics while merely 3 per cent use prescriptive analytics. Therefore, teams do not have the right type of insights that they require to make data-driven decisions.

The second issue is that even when they are clear about what data they need, the teams are not able to access all this data available to them effectively. Here the organizational silos play a critical role in sub-optimal value-creation from data. Traditionally, teams are used to working in silos, and are often not open to sharing information with other functions. This is either due to the traditional mindset of hoarding information or they work with legacy systems that restrict easy sharing of data across functions. This is

the key reason why teams lack the end-to-end view critical to drive effective data-driven decision-making. It's no surprise that 97 per cent of executives think data silos have a negative effect on business.⁶

Another culprit for inadequate access to data is the sub-optimal approach to data security that most organizations adopt, which is common standards of security across organizations irrespective of the role and requirements. If you take a 'one size fits all' approach to data security, you will end up sharing either too little or too much. Typically, in large and/or legacy organizations, former is the case.

Issue 2: Teams don't have the data they need



The third issue arises because of the dynamic nature of business problems today. As the business problems keep evolving, so do the data requirements and this creates a lag effect, where decision-makers do not get the data at the right time. This is because of the long data management cycles making it difficult to incorporate iterations that quickly. During the design phase, data pipelines are created with the bigger picture in mind. But when used for decision-making, teams often realize they need either more detailed or different information. These evolving data requirements are difficult

to address owing to tedious data management cycles and system design constraints.

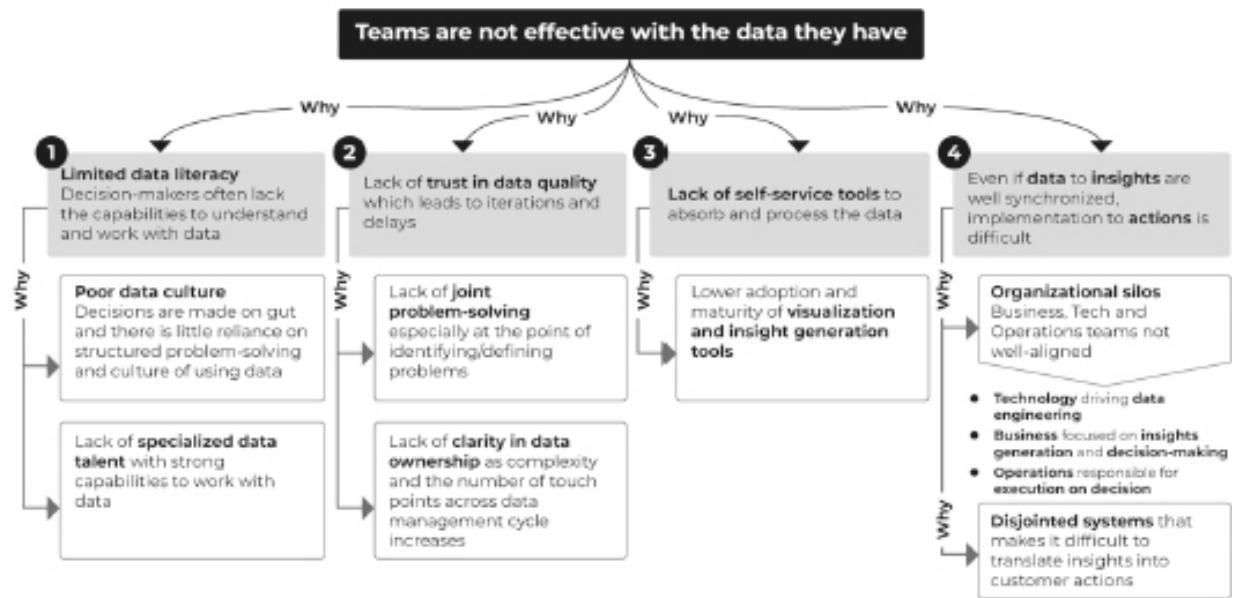
Symptom 3: Teams are not effective with the data they have

One of the core reasons why teams are not effective with the data they have is that they are unable to effectively leverage data to take decisions and drive actions. But why is that the case when data is available to them in so much abundance. It is because most of them come from a non-technical background with limited knowledge or know-how to use data or even understand the data that is available to them. As a result, they either use the data in an inaccurate manner or do not use it at all, instead relying on their gut or experience to make critical decisions. This lack of capabilities and methods to use data is evidence of a lack of data culture. And despite the data explosion happening for a number of years, it is still a deep-rooted issue in many organizations. It's the way organizations have been working for decades together and therefore requires a deeper change in the organization's DNA. In fact, data culture has been identified as the numero uno of all the challenges in realizing full potential from data and analytics.⁷ In addition to that, as I highlighted in symptom 2 as well, there is a skill gap in terms of the high-quality data talent that comes with the right set of skills which go beyond data expertise to problem-solving, storytelling and deeper domain expertise. This is why the insights generated are often either disconnected to the business problem or are not impactful enough to be leveraged for data-driven decision-making and driving desired actions.

Another reason for not being effective with data is the fact that they do not trust the data because they have not been a part of the entire data process. The key reason behind that is that businesses, technology, and the operations team working in silos, are not jointly driving the problem identification process. It is largely led by the data or tech teams, who often lack the business perspective.

Additionally, owing to the complexity of the data ecosystem today, there is no clear ownership of data. Due to the growing number of touchpoints, multiple stakeholders across the value chain contribute to the data sets, at varying speed and scale. As a result, no single team can control or maintain the data quality, adding to any teams' apprehension to use it with confidence.

Issue 3: Teams are not effective with the data they have



The third reason is the lack of effective self-service tools that enable easy and seamless absorption and processing of data. These self-service tools enable the end user to interact with the data on their own and work with it as per their specific requirements, which is a critical factor in making the most effective use of data. This is because of the low level of adoption of insight generation and visualization tools by organizations. The average adoption rate of business intelligence in a mid to large organization is around 15 per cent only.⁸

And last, but by no means the least, even when the data and insight generation are well synchronized, translating them into actions is very difficult. This is because the process of translating insights to action is not done in a collaborative manner across

functions. As organizational silos exist, each function or team has independent responsibilities and are not working together. Therefore, it becomes very difficult to effectively translate the insights into actions. While the technology team drives the data infrastructure, the business is responsible for insight generation based on the business problem they are facing while the operations team is responsible for executing the decisions.

Furthermore, the disjointed systems also make it difficult to effectively translate insights into actions. Most organizations are working with systems that are not well-integrated or connected with each other, because of which translation of data or insights from one system to another to drive action is not possible, again making implementation a difficult task to achieve.

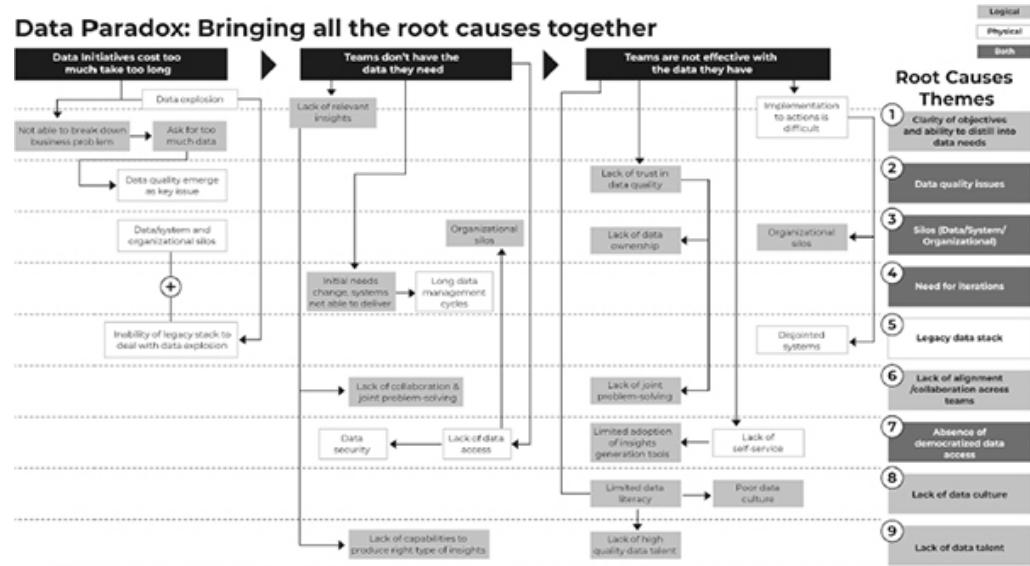
Bringing it all together: Nine root causes

When you look closely at all the issues that I have highlighted under each of the three broad symptoms, you will notice two things. One is that many of them overlap in some form or the other across the three symptoms, like alignment between IT, business and operations. And that the nature of these issues has some common characteristics. For example, not being able to break down a business problem is a purely 'logical' issue, while lack of self-service tools is a 'physical' issue. And then there is another category, like data-quality issue which has aspects of both physical and logical issues. So, when I bring all these root causes together, I see some common themes or root causes emerge. Following are nine root causes that emerge:

1. Clarity of objectives and ability to distil into data needs:

needs: Since stakeholders are unable to clearly define the business problem and are not able to narrow down their data requirements, they end up asking for too much data. Additionally, they lack relevant insights because what they get is not what they want. Working in silos, the data teams often do not work in collaboration with

the business owners to define the business problems properly. And most of the insights generated are still of descriptive nature, while the business owners need the why and how to take appropriate actions. Therefore, perhaps the biggest root cause of the Data Paradox is lack of clarity on the business problems which result in an inability to refine the data requirements.



2. Data-quality issues: The fact that there is too much data makes it difficult to monitor and manage the quality of such large and dynamic data sets. And since teams are not sure of how the insights were arrived at, and the quality of data that has been used to derive these insights that are made available to them, they are reluctant to use it. Therefore, the second biggest root cause theme is the data-quality concerns that lead to sub-optimal utilization of data.

3. Silos (data/system/organizational): Whether it is the system, the team or the data itself, multiple silos exist in an organization that contribute to sub-optimal value generation from data. Because of these silos, the teams do not have access to all the data that they need, and they are also reluctant to use it because of unclear data ownership. And these silos also cause bottlenecks in

teams collaborating across the organization to make effective use of data, especially when it comes to translating insights into actions.

4. **Need for iterations:** As businesses are increasingly operating in a dynamic environment and the data is pouring in high volume, variety and velocity, the nature of business problems also keeps evolving and as a result the data requirements also keep changing. The need for iteration is high and very often, which is not manageable due to long data management cycles. This is because legacy systems are not capable of handling Big Data and translating data into insight at speed and in real time. On top of that, an even bigger problem is that the complex legacy processes operate with the mindset of getting it 'right the first time'. It typically means taking more time to design, analyse and get it right the first time. This makes it difficult to drive iterative, experimentation-based processes.
5. **Legacy data stack:** Another root cause theme that emerges is the constraints due to legacy infrastructure that most organizations are struggling to upgrade or integrate with new systems and processes. This leads to higher costs and more time spent on managing and integrating them. This legacy infrastructure has grown inorganically by cobbling up disparate systems over time, often sourced from different organizations. The main limitation is their inability to leverage Big Data, most of which is unstructured, to drive insights. This legacy infrastructure lacks the ability to achieve scalability, flexibility and granularity. This root cause is often well understood by many organizations, and is the first thing they try to address, but they lack a systematic approach to handle it effectively.
6. **Lack of alignment/collaboration across teams:** Lack of collaboration between the business, IT and operations teams results in data strategies that are not

well defined. As all three teams are not working jointly, the lack of collaboration has repercussions across the data cycle starting with problem identification to insight generation to implementation of action. This is because the teams often are unable to agree on the insights derived, or even if they do, the organizational silos make it difficult to implement them jointly. This, along with root cause 1, is perhaps one of the most important root causes for organizations being stuck in a data paradox.

7. **Absence of democratized data access:** Traditionally, organizations work in silos and teams are not open to sharing data with other functions. Therefore, teams lack an end-to-end view on data, which adversely affects the decision-making process. Additionally, though data security is a critical issue in the Big Data world, without a balanced approach it can easily become a bottleneck for seamless access to data to various decision-makers across the value chain for effective decision-making. And even if the data is made available to business owners, it is not easy to translate into insights customized as per business owners' requirements owing to limited adoption of self-service tools. These challenges contribute to lack of data democratization—making the right data available to the right people at the right time.
8. **Lack of data culture:** Even when data is available in abundance and organizations are continuing to implement tools and systems to generate meaningful insights, teams are not able to use it. Owing to lack of data literacy, most business owners who come from a non-technical background lack data analysis and interpretation capabilities to leverage data effectively. Most organizations still have not implemented an organization-wide use of a structured data-driven decision-making approach and rely more on gut-based approach to problem-solving. This indicates the lack of a data-driven culture.

9. Lack of high-quality data talent: And lastly, another key reason for teams not having the right data as well as not being effective with it stems from the fact that the kind of data talent required to solve any business problem end-to-end is scarce. Many data experts may possess good data-handling and analysis skills, but lack problem-solving, storytelling and deeper business domain expertise. As a result, there is a disconnect between the data, the business and IT, as data experts are unable to contribute to solving a problem end to end which is critical to drive desired impact. This significantly hinders the collaborative efforts between business, IT and operations. This is why the insights generated are either not up to the mark or are completely disconnected to the business problem at hand.

Now if you look at all the issues categorized as logical, physical or both, a significant observation emerges, which I pointed out in the beginning of the chapter—the issues that organizations face, are more logical than physical ones. Of all the root cause themes that have emerged from this exercise, the majority are either logical or involve some aspect of them. Physical issues are fewer in comparison. Now imagine putting in all your time and effort to solve only the physical issues, while the others are just growing in scale and complexity. So, the real question is not whether you will end up in a paradoxical state or not, but when.

Key takeaways

- Organizations spend a disproportionate amount of time and money on solving the ‘physical’ issues to handle the Data Paradox but get stuck in a vicious cycle of adding to the data and tech infrastructure without effectively solving the business problems.

- Deep diving into the symptoms reveal nine underlying causes, most of which are logical in nature or have some aspect of 'logical' in them. Logical issues being unclear business problems and inability to break it down into data requirements.
- Organizations must recalibrate their approach and focus a lot more on solving 'logical' issues, or they will never be able to conquer the data paradox and thrive in the data-first world.



SECTION II

MAXIMIZING VALUE IN THE DATA-FIRST WORLD

'Though this be madness, yet there is method in it.'

—Hamlet (*Act 2, Scene 2*), William Shakespeare

A Unified Solution Framework

Thirteen Mantras for Data Success

'There's no use talking about the problem unless you talk about the solution.'

—Betty Williams,
Peace activist and Nobel Peace Prize co-recipient

Introduction

In the previous section, I discussed the exceptional growth of data over the past few years. How it has transformed every aspect of human life and created tremendous opportunities for organizations to generate value. However, it is undeniable that there is a big gap between the potential and the value realized from data by enterprises. This is because they are stuck in the Data Paradox, where on the one hand they face a data deluge which they are unable to deal with, while on the other, they face an insights and decisions drought. And this is exactly the situation best described by the quote, 'water, water everywhere, nor any drop to drink'—surrounded by an ocean of data, but always parched for insights.

When we dug a little deeper into the issues, we unearthed a very significant pattern, a trap that organizations are falling into. Most organizations, when undertaking data initiatives, are investing a large part of their time and money in solving the physical issues like adding technology infrastructure to capture all the data that is pouring in, while paying less attention to solving the logical issues like defining the business problems clearly to narrow down the data requirements. This disproportionate focus is evolving into a

quicksand situation where, the more the organizations try to dig themselves out, the deeper they get pulled into it. The deluge gets bigger and bigger as more and more data keeps pouring in, while the insights drought remains severe and widespread in the organization.

So, how can organizations break the pattern to effectively deal with the Data Paradox and maximize value from data to succeed in the data-first world?

'Framework' to the rescue

When you look at the Data Paradox, at first glance, it looks like a complex problem that looks almost impossible to solve. Yes, data will keep growing, growing at a pace that seems impossible to manage. Organizations that are set in their ways, will continue to show resistance to change. And when you look at the root causes, there are some that need a drastic change in the way organizations are used to operating for years. It seems like a daunting task, and one might wonder, can this really be done?

Well, to that, my answer is yes, absolutely. But we cannot use the same old traditional, time-worn approach to solving an issue of such unprecedented nature. New problems need new solutions. What we need is an innovative and practical approach. In my experience, whenever one is faced by a complex problem that doesn't seem to have an obvious or a straightforward solution, the best way to approach it is to break it down into manageable pieces. A 'framework' is the best way to break a problem down in a logical manner so that every component solves for a specific problem or set of problems, and when put together all the components become a comprehensive guide to solving the problem at hand.

With that, let me introduce my thirteen-component solution framework that will enable companies to maximize value from data. It provides a holistic view on how an organization can win in the data-first world and be prepared for the AI age.

Unified Solution Framework (USF)

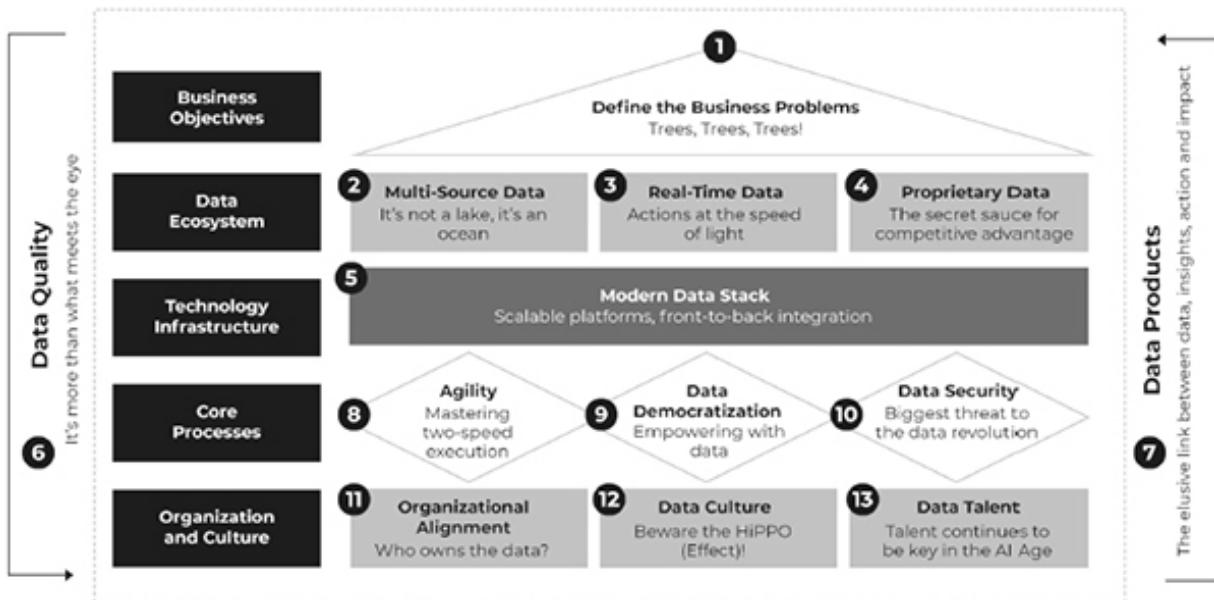
In the Big Data world, where data is available in abundance, it is important to understand not just what to leverage, but also how to leverage to maximize returns. The USF is my way of logically putting all the components together that play a critical role in leveraging data. This framework emphasizes the importance and role of each component in enabling data success, and its relationship with other components that highlights its part in the overall effectiveness of a data initiative.

The starting point of data initiatives for an organization could either be a business problem or data infrastructure build-out, but irrespective of what the starting point is, this framework can prove to be a very useful blueprint that can help them navigate their data initiatives more effectively. Each one of these components is a necessary building block to build a data-driven organization to succeed in the data-first world and the AI age.

So, let me now introduce you to the thirteen characters (read: components) of my story, to which the rest of the chapters in Section II will be dedicated.

Whether it is about starting a data initiative or building a data-driven organization, the starting point for any organization should be to define the **Business Objectives**. Defining the business objectives is like asking the 'why'. Why are we doing it? Asking why, is the starting point and critical to clearly define the problem that we are trying to solve. Which brings me to the first component of the framework, from where it all begins, making it the most critical component of the framework—'defining the business problem'.

The Unified Solution Framework (USF): Thirteen key components



1. Define the business problems: *Trees, trees, trees*

Someone rightly said, 'If you torture data long enough, it will confess to anything.' But this is not helpful, because it might lead to wrong conclusions, errors in judgement and disastrous business outcomes. It is therefore critical for organizations to invest time and effort upfront, to better define and narrow down the business problem before they even start looking at data. Once you narrow the business problem, the next step is to identify the data required to solve the problem. Accurately defining the data requirements can significantly reduce the data requirements and make the data pipelines more manageable, automatically bringing down the complexity of working with Big Data.

Building 'trees', especially KPI (key performance indicator) trees, is the most effective way to both narrow down the business problems and define the data requirements. Through KPI trees, the business problem can be broken down into multiple yet manageable chunks to identify the key drivers that are expected to affect the outcome. The data requirements specific to these drivers can then be identified and data pipelines can be built accordingly, for effective problem-solving.

Once the business problem is clearly defined, and the data requirements are identified, we then ask the question ‘what?’ What will help us achieve these business objectives? And the answer lies in the extensive data ecosystem that is made available to us owing to the explosive growth that data has witnessed over the past few years across the 3V dimensions. The data ecosystem must be sifted through very carefully to identify the most effective combination of data sets, which would enable us to generate deeper and more meaningful insights relevant to the business problem at hand, without being overwhelmed by it in the long run.

The Data Ecosystem consists of both external and internal data sets, which are either structured or unstructured, made available to us in batches or streaming in real time. From this vast data ecosystem that is available to us in the digital world, I have highlighted three major types of data sets that, in my experience, drive disproportionate value generation for organizations. Let me introduce these three to you, which form the next three components of the USF.

2. Multi-source data: *It's not a lake, it's an ocean*

In the Big Data world, most of the data is available outside the organization today. And as data is being captured through newer touchpoints, the sources for it continue to expand in variety. Moreover, as organizations operate as a part of a bigger ecosystem, it is no longer enough to rely on internal data only. Thus, it is becoming increasingly important to use external data along with internal data for decision-making.

Integrating two or more data sources, external, internal or both, enables organizations to derive much higher value from the data and significantly enhances the depth and quality of insights generated. Having said that, integrating data obtained from multiple sources has become fairly complex due to increased variety and complexity of data. Traditional approach of setting up an operational data store doesn't work in the Big Data world, hence organizations need to move towards newer approaches like data mesh and data fabric.

3. Real-time data: *Actions at the speed of light*

In the digital age, the ability to respond quickly in real time or near real time is key for businesses to win. And one of the biggest shifts brought about by digital is the availability of data in real time, which enables organizations to do just that. The digital world generates data not just in high volume, but at high velocity as well. Real-time data is pouring in continuously, an aspect that is difficult to manage, because the treatment of real-time data at every stage of the data-management cycle needs to be instantaneous. Newer and more efficient technologies that are available today have helped organizations build the right technology infrastructure to support such a high-velocity data management value chain.

But not every business problem or use case requires use of real-time data and storing and analysing real-time data in a timely manner is an arduous and expensive exercise. So, organizations must identify the right use cases that are ideal for using real-time data by analysing the impact vs effort to narrow down the right ones.

4. Proprietary data: *The secret sauce for competitive advantage*

As I pointed out before, most of the data is available outside the organization, openly available to everyone to leverage. What can really create differentiation is the tacit knowledge available within the organization at individual, team or function level, which is codified to create proprietary knowledge. The knowledge that is generated and captured over time, through experiences, accumulated knowledge and evidence collected through the years, is unique for an organization. Capturing this tacit knowledge and codifying it, to be used by the organization in a repeatable manner is the key to create a competitive advantage in the data-first world. It is like that pinch of salt that immediately enhances the taste of the food, making it perfectly delectable. To develop and leverage proprietary data, organizations must create a 'knowledge cycle' that has two components—knowledge creators who are

sources of tacit knowledge and knowledge seekers who learn from the codified tacit knowledge.

While these do not form an exhaustive list of data sources, in my opinion, these are the three main sources of opportunities of the Big Data world that organizations must leverage to maximize value from data.

So you have identified the business problem, and have all these multiple types of data pouring in. Now what? Now, you need a way to capture, store and translate all this data into a format that is easy to access, understand and consume. How would you do that? For that you need a foundation, a base that provides a solid footing for the organization.

Technology infrastructure, and specifically the **Modern Data Stack**, is the necessary foundation and historically the most important aspect of any data initiative. It is the fifth component of the USF.

5. Modern data stack: *Scalable platforms, front-to-back integration*

As I highlighted before in Section I, the legacy data stacks that many traditional organizations have are unable to effectively consume, integrate or process the increasingly complex and ever-expanding data, resulting in costly and time-consuming data initiatives. Modern data stack consists of loosely coupled yet tightly connected layers of the data stack, hosted on the cloud, to enable rapid scalability, accessibility and adaptability.

There are seven critical technology shifts that organizations must embrace to build a modern data stack. The most significant is the move from on-premises (on prem) infrastructure to cloud, largely enabled by the hyperscalers (large cloud service providers that provide storage, computing and more), enabling flexibility and elasticity requirements to scale on demand. Organizations have to move from batch to stream-processing using real-time infrastructure which enables them to respond to rapidly changing business requirements. Furthermore, monolithic architectures

must be replaced with the flexible yet interconnected structure of microservices based architecture, where a collection of small, independent services can be developed, deployed and scaled independently. A hybrid approach with capabilities of both a data warehouse and a data lake is helpful in meeting the needs of businesses in the Big Data world. Data warehouses enable storing data in structured form for BI and reporting. Data lakes are designed to store all types of data structured, or unstructured, at scale.

Furthermore, organizations need to move from a technology-driven approach of bringing all the data together to a logical approach of integration depending on the specific use cases or business requirements. Organizations must move from AI experimentation to become AI first, by leveraging the transformative value of Gen AI through a data stack equipped with enhanced capabilities across the data management cycle. And finally, the consumption layer of the data stack has to be modernized with self-serve analytics to enable the business owners and users to interact with the data as and when they need and generate insights as per their requirements.

Before I move on to the next layer, let me talk about the two **integrators** that vertically cut across all the layers of the USF and act as a link that brings together all the layers of the framework.

6. Data quality: *It's more than what meets the eye*

Data quality is an essential component to any data initiative, which is relevant across all the layers of the USF, impacting every other component. Getting it right from the word go at each stage is very critical to data success. The notion that Big Data compensates for quality issues, doesn't hold true in reality, especially when decisions or actions are needed at an individual level. For this, data has to be very precise and error-free at a granular level.

This is where Big Data adds to the complexity of managing quality owing to the increased number of sources and variety of data pipelines where quality issues can arise in multiple ways.

Moreover, the traditional dimensions being used to measure and ensure data quality are not enough. The purview of data quality is expanding, becoming bigger and more fluid. Data quality requires a context-first approach, determined by the intended use and the criticality of the business problem at hand.

7. Data products: *The elusive link between data, action and impact*

The most effective solution to solving the Data Paradox is 'productization' of data. Data products are digital assets built by integrating various elements across the data stack to deliver specific outcomes on the DIAI framework in a repeatable manner. These solutions are built to accelerate the data management cycle because they are readily available to the business owners to plug and play as and when required.

Data products are the most effective way to significantly reduce the value leakage at every stage of the DIAI framework, enabling organizations to maximize value from data. These can be of varied capabilities and complexity depending upon intended outcomes. They are built by identifying repeatable, commonly used data assets across various projects and converting them into products that can then be used like Lego blocks. These data products are also constantly evolved and iterated upon based on adoption and feedback from business users and fed with more and more data as and when it is captured. With growing advancements in technology and evolving requirements of the business, I foresee a surge in adoption of more advanced data products—AI-powered, real-time, industry-tailored and more.

Now going back to the horizontal layers of the USF, once the foundation is all set for the data to be leveraged in a most effective manner with the Modern Data Stack, we need **core processes** to operationalize this data stack to drive adoption by making it accessible, available and easy to use for decision-makers. The Core Processes consist of three components:

8. Agility: Master two-speed execution

There is so much data that organizations are dealing with that most data initiatives tend to become very big and take too much time. But businesses are operating in a dynamic environment. So, while organizations are trying to solve a business problem, more often than not, the business requirement itself changes and so does the nature of the underlying data required. So, organizations are at risk where the longer a data initiative takes, higher the chances are of it becoming less relevant.

To avoid such a situation, businesses need to approach data initiatives such that they are able to respond to changing business needs quickly, while building long-term capabilities in parallel. Traditional Big Bang approaches to data initiatives will not serve the purpose. To achieve such speed and flexibility, organizations must adopt a two-speed approach. Speed One is where you identify a high-impact use case which is solving a specific business problem and tackle that first. And since it is solving a specific business problem, it would typically take less time and enable organizations to learn. Speed Two is when you bring the speed One use cases together and build long-term capabilities. As you continue to connect Speed One use cases, the underlying data infrastructure grows, you would reach an inflection point where it can become a source of uncovering new insights.

9. Data democratization: Empowering with data

While historically data has been seen as a source of power and, hence, people have been hesitant to share it, in the Big Data world, where data is available in abundance and mostly freely, it is impossible for any one person to hoard it. It is therefore a great opportunity for organizations to break out of the traditional paradigm of selective or limited access to data, to make it freely and easily accessible to decision-makers across the value chain, empowering them with it to make informed data-driven decisions.

To do so successfully, the organizations must first understand the data landscape to identify data silos created either due to lack of collaborative mindset or the legacy processes and systems.

Breaking down these data silos by bringing data together and making it available and easily accessible to all business users is the first step to enable data democratization. In addition, they must implement self-serve capabilities that would boost the adoption of data-driven decision-making. Self-serve tools enable executives to easily access the data, perform some basic analyses on it and create data visualizations, dashboards and reports as per their business requirements. This entire process is iterative and needs continuous monitoring, evaluation and upgrades based on changing business needs and decision-maker's requirements that keep evolving over time.

10. Data security: *Biggest threat for the data revolution*

In the Big Data world, and especially as the AI age dawns upon us, the security threats are becoming much bigger, more frequent and highly sophisticated. The risk and the costs associated with data security has gone up exponentially too, which requires organizations to invest in data security continuously.

Today, data is pouring in from all directions, and the number of touchpoints or endpoints in an organization is way too high and scattered owing to the users, data and resources being spread across the globe. The traditional approach of 'castle and moat' does not work any longer, so organizations have to move towards zero-trust framework—prioritizing strict identity authentication for every individual, device or application seeking access to the network resources. But such zero-trust policy raises an inherent tension between data security and democratization. Therefore, building a zero-trust architecture requires a context-driven approach. Because 'one size fits all', will likely lead to sub-optimal access. The context should be built based on three key considerations. First, the criticality and sensitivity of data. Second, data access is customized as per user profile or persona—their roles, levels or responsibilities, etc. And lastly, the pattern of activity, where behaviour-based security is provided by continuously monitoring all activities against anomalies and inconsistencies. AI is beginning to play a critical role in automating

the data security value chain enabling organizations to protect, detect, investigate and respond to security threats with greater speed and precision.

And this brings me to the final layer of my USF, the **organization and culture**, which may be the 'softer' layer of the solution framework but has a permeating impact across the organization. Developing a positive data culture, driving a data-driven organizational alignment and nurturing specialized data talent are key to initiating a long-term change in the organization's DNA:

11. **Organizational alignment: Who owns the data?**

Traditionally, organizations have followed a functional approach to data ownership, where functions were responsible for decision-making and for data for their own functions. But in the Big Data world, where multiple stakeholders both consume as well as contribute to the data, it is not easy to clearly assign ownership or accountability of data to one single person or team which makes data governance challenging. This lack of clarity on data ownership makes it difficult to ensure the quality of data that keeps pouring in from all directions. Additionally, conventional organizational structures prevent effective collaboration in framing business issues collectively, resulting in wasted time and effort in producing ineffective or irrelevant insights.

The success of any data initiative depends on three key considerations: organization structure, decision-making and execution roles. Organizations have to make a strategic choice between horizontal and vertical alignment, and the level of decentralization in their decision-making. In addition, they must address the inherent tension between ownership and collaboration in the Big Data world. Adopting a data product-centric approach can become a potential solution for organizations looking to effectively align teams with clear end-to-end ownership of the product and the underlying data and enable collaboration between teams with varied expertise to work together to deliver the target outcomes.

12. Data culture: *Beware the HiPPO (effect)*!

Despite all investments in data infrastructure, a data-driven organization is impossible to build unless people at every level of the organization are data-driven as well. Traditionally, organizations and the executives who run them have been relying on their gut and experience of the HiPPO—Highest Paid Person's Opinion—to run businesses. In the Big Data world, where data is exploding by the day, relying just on one's gut is no longer good enough. This not only brings a high risk of errors in decisions, does not encourage collective accountability, and nor does it benefit from the vast amount of data that is available.

Organizations must consciously move away from HiPPO culture to a data-driven decision-making process. There are three key levers to achieve this. Organizations must invest in enhancing data literacy across all levels by building data analysis, interpretation and storytelling capabilities. Second, making data available to all decision-makers and establishing the right processes to promote effective use of data. And lastly, it requires a push from the HiPPO himself/herself, where the leaders lead by example and prioritize data in their decision-making to inspire and influence employees at all levels to do the same.

13. Data talent: *Talent continues to be key in the AI age!*

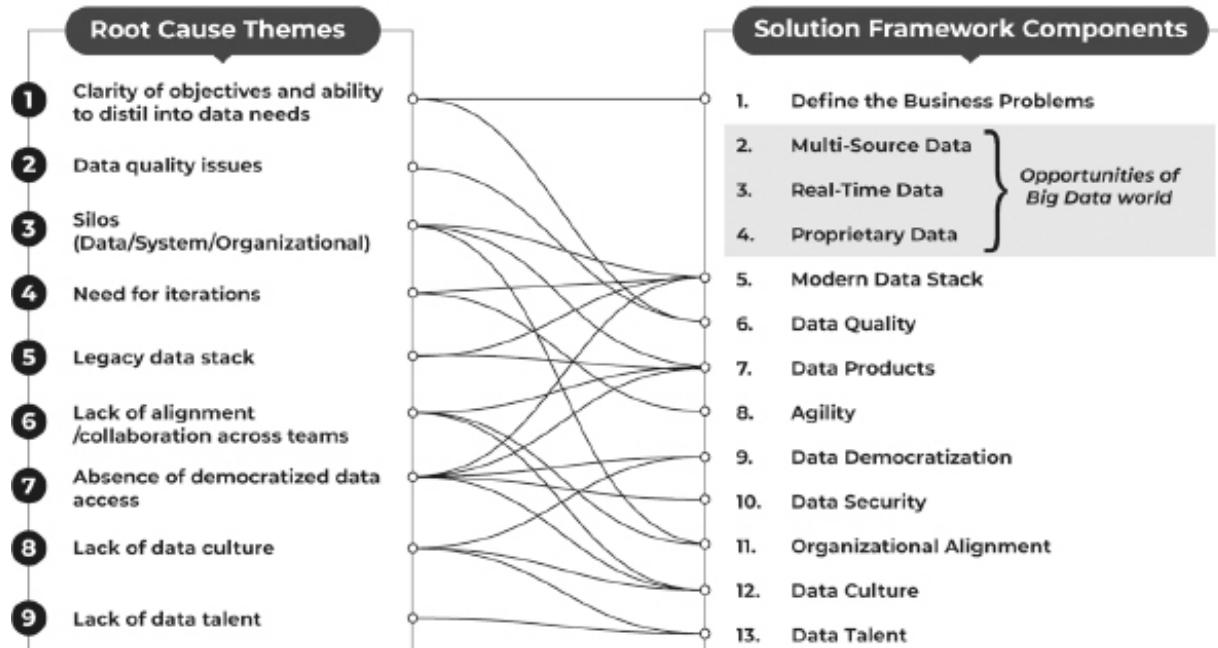
As the data landscape continues to become more and more complex and the AI age unfolds, the importance of building the right data talent becomes even more critical to unlock the full potential from data. Many specialized data roles are required to come together to deliver on data initiatives including data scientists, data engineers, business analysts, data architects and more. However, I foresee this data talent landscape rapidly evolving. In addition to core data skills, data professionals would also require a deeper understanding of the business domain to be able to solve a business problem end to end. As AI technologies get adopted at scale in organizations, the data-science roles are likely to shrink as bulk of the model building will likely be done by machines. What would become more important is problem-solving

skills and within that, problem-identification where domain knowledge would be important. Another important skill would be their creativity in storytelling. On the other end of the spectrum, the role of data engineers would become more complex as they have to manage massive amounts of data which is the foundation for AI. Similarly, data architects will play an even more pivotal role in shaping and optimizing data infrastructures to support the data-driven strategies and AI initiatives of organizations.

USF: The guide to solving the root causes of Data Paradox

All nine root causes for the Data Paradox identified in Chapter 5, *The Root Cause*, can be solved by leveraging the thirteen components that I have just introduced. There is a ‘many-to-many’ mapping between the root causes and the solution components. Let’s see quickly how that works:

The Unified Solution Framework solves for the root causes while exploiting the data opportunity



Data, which is the subject of this book, is the **opportunity** made available to us by the Big Data world. And multi-source data, real-time data and proprietary data (component 2, 3 and 4) are the most important aspects of the data ecosystem. These components are there in the framework not because they solve a problem but because they have the potential to add disproportionate value for organizations. Rest of the components of the USF are about addressing one or many of the root causes of the Data Paradox that organizations face.

The first root cause we identified was the **lack of clarity in business objectives and distilling the data needs**. It can be addressed by the first component of the USF: defining the business problem. Breaking down the business problem into manageable pieces and narrowing down the data needs makes the problem easier to solve and helps eliminate wastage of storage and computing power.

The second root cause, **data quality issues** can be tackled by component 6, data quality. Addressing the data quality at every layer of the USF, and evaluating quality based on the context it is being used in, will help organizations address the data quality issues effectively in the Big Data world.

The third root cause identified are **silos** of various kinds, whether it is data, system or organizational, that create bottlenecks in realizing full potential with data. It can be addressed by building a modern data stack that brings all the organization's data into one place in a consumable format, building data products that can be used in a repeatable manner by multiple teams and adopting an innovative organizational alignment around data products to create a balance between ownership and collaboration.

The fourth one, **need for iteration** can be enabled through modern data stack and agility that enables organizations to incorporate iterations quickly and respond to changing data needs faster.

The shortcomings of **legacy data stack**, which is the fifth root cause, can also be addressed by building a more scalable, adaptive

and agile data stack and building business solutions that can be used in a repeatable manner across teams and use cases.

The sixth one, **lack of alignment/collaboration across teams** can be effectively addressed through data products and aligning teams based on data product-centric approach that could ease the tension between ownership and collaboration.

Absence of democratized data access is the seventh one. It can be proactively addressed by democratizing access and use of data across the organization, building the right level of data security while keeping democratization in mind, and building a data-driven culture that encourages and facilitates data sharing.

The eighth root cause, **lack of data culture**, can be addressed by democratizing use of data and encouraging and facilitating data literacy, and role-modelling and rewarding a data-driven decision-making process at every level in the organization.

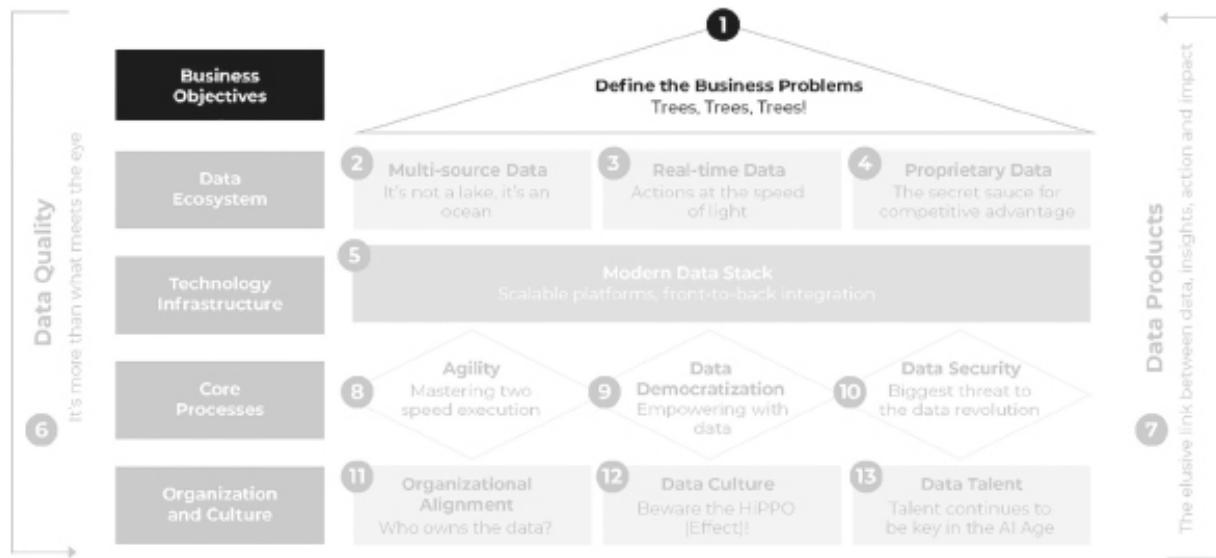
And the ninth one, **lack of data talent**, can be addressed by establishing the right data culture and prioritizing problem-solving and business domain knowledge along with core data skills when hiring specialized data talent.

The Unified Solution Framework is an innovative and practical guide which will help organizations not only tackle the many challenges in working with data more effectively but also achieve transformational value with data, critical to succeed in the data-first world. And while this framework provides a much-needed direction to approach data problems, it is still something that has not been fully mastered yet and so organizations must also be open to learning and course-correcting as they go about it.

In the next few chapters, I will talk about each of these components in detail and highlight their role in solving the Data Paradox effectively.

LAYER 1

BUSINESS OBJECTIVES



'Asking the why?'

Define the Business Problems

Trees, Trees, Trees!

'A problem well-stated is a problem half-solved.'

*—John Dewey,
Philosopher, psychologist and education reformer*

As explained in Section I, organizations usually tackle business problems by accumulating extensive data, assuming that more data yields better insights. But as they say, 'A ship that sails without a compass is bound to get lost at sea.' In the Big Data world, where data is exploding in all directions, this approach generally translates into dealing with a sea of data, abundant and overwhelming. As a result, organizations end up spending significant amounts of money, time and resources on gathering and analysing large amounts of data in hopes to find answers or breakthroughs that, more often, do not happen. Often, they get caught up in the process and get misdirected or lost.

In this era of continuous technology advancements, exceptional talent and ample data, often organizations start a data initiative with a lot of enthusiasm, only to realize at a later stage that it is not solving the problem they set out to solve or they are solving the wrong problem. Why does that happen? Because we are programmed to jump right into solutioning mode, without spending enough time understanding and defining the problem. In my experience, it is this exercise of defining the business problem that organizations must dedicate maximum amount of time, effort and resources on. As Albert Einstein once said, 'If I were given one hour

to save the planet, I would spend 59 minutes defining the problem and 1 minute resolving it.'

Let's explore in this chapter how we can do this better and get our data initiatives to the right start.

Identifying the 'bottleneck'

Effective problem-solving hinges on identifying and tackling the bottleneck or the weakest link. A concept that is also the basis of one of my all-time favourite books, *The Goal* by Eli Goldratt. A book that I recommend everyone should read. According to his 'theory of constraints', and I paraphrase, every process (he takes the example of manufacturing) has some bottlenecks; the weak links, whose capacity is equal to or less than the demand from it. In other words, these bottlenecks are the reason for the hold-up and must be addressed. Because any improvement made upstream would result in additional inventory in the system and improvement downstream would result in idle capacity. The key to effective problem-solving lies in identifying these bottlenecks and managing those effectively to improve the overall throughput of the process. So, narrowing down on the bottleneck in any situation and then focusing your solutioning efforts around that core issue is the key to making progress on any problem.



Problem-solving needs a structured approach

Because of these insights, I could easily relate to the problem-solving techniques I saw at McKinsey, a fact-based, structured approach to problem-solving ingrained in every aspect of its business.¹ Beyond McKinsey as well, through all these years, I have always relied on and recommended adopting a structured approach to problem-solving. During my experience of leading large transformation initiatives, whether it was at Fidelity, or Flipkart, or the Fortune 500 companies that we serve at Incedo, I have found that structured problem-solving approach is relevant for all organizations, irrespective of the size, nature or industry that they are operating in, even more so in the data-first world.

While structured problem-solving could mean different things for different people, the basic approach remains the same—start by bringing clarity on the problem to be solved. For this, understand the context, which includes the stakeholders, their objectives and the challenges they are facing. Having narrowed the problem, the next step is to break it down into actionable components and then to prioritize them. This significantly reduces the complexity and ambiguity around the problem(s) to be solved, making the whole process more effective and easier to drive. In the Big Data world, narrowing it down on a use-case basis, to make the data requirements and therefore the overall initiative more manageable is critical to the success of any data initiative.

So, any complex problem can be reduced to a group of smaller, simpler problems that can be solved individually. Simple, is it not? It is really surprising how often people skip this step, or do not invest enough time on it. Rather, they instinctively get into the problem-solving mode, carrying the biases of their own assumptions and judgements. Let's explore this with a very typical data-intensive business problem: creating a single view of the customer!

The challenge of creating a 'single view of customer'

A growing demand for Netflixization of content and Amazonification of services prompted organizations to seek a consolidated view of their customers and operations to better engage with them and deliver personalized services and offerings to them. So, it became critical for organizations to build what they call 'single view of customer' or '360-degree customer view'—an aggregated, consistent and holistic representation of the data held by an organization about its customers that can be viewed in one place, such as a single page.

The question was, how could they do so?

In the past decade, Google had emerged as a pioneer in the use of data and was successful in consolidating its customer data and building analytical models on its users and their activities, to provide innovative services and customized recommendations. Many large enterprises believed that by replicating this Google model, they would achieve similar success. They attempted to bring all the data on their customers together at one place, in the hope of developing similar innovative and customized services.

This effort to build a centralized repository for all the data within the organization, which can be drawn upon when required, led them to 'single source of truth' (SSOT), which sounded like the perfect solution to the data silos. Pretty much every organization was ready to jump on the bandwagon. However, not many have been successful. Despite spending huge amounts of money, time and resources to build a SSOT, a recent survey reports that so far only 14 per cent of organizations have been successful in building a 360-degree customer view while a whopping 82 per cent of them still aspire to achieve that.²

Why? Because there is an inherent flaw in this theory that most organizations missed and they got caught up in the whirlwinds of the Big Data world. They underestimated the pace at which data is growing and expanding and did not take into account the fact that most of the data today is generated and consumed in real time. So,

bringing it all together became a mammoth and unending exercise, and especially ensuring data quality across a vast variety of data sources became a particularly intractable problem. While enterprises were busy collecting all the data and bringing it together, opportunities were still being missed, revenues were still being lost. This is why SSOT, while an ideal concept, is very difficult to achieve.

Sure, conventional wisdom has it that all the data points brought together in a consistent way at a centralized location would make it more effective for decision-makers to access and draw insights from. But going all out to capture every datapoint as part of building the SSOT, while well-intended, can become impractical and in most cases, not deliver the desired benefits.

It is exactly like the situation that I faced at Flipkart that I shared in the introduction. The efforts to build a single source of truth got so big and spun out of control so quickly that it eventually resulted in gridlocking the whole organization. A dream project to bring all that data on millions of customers and thousands of operational metrics and dealing with all that data together ended up becoming a problem rather than a solution!

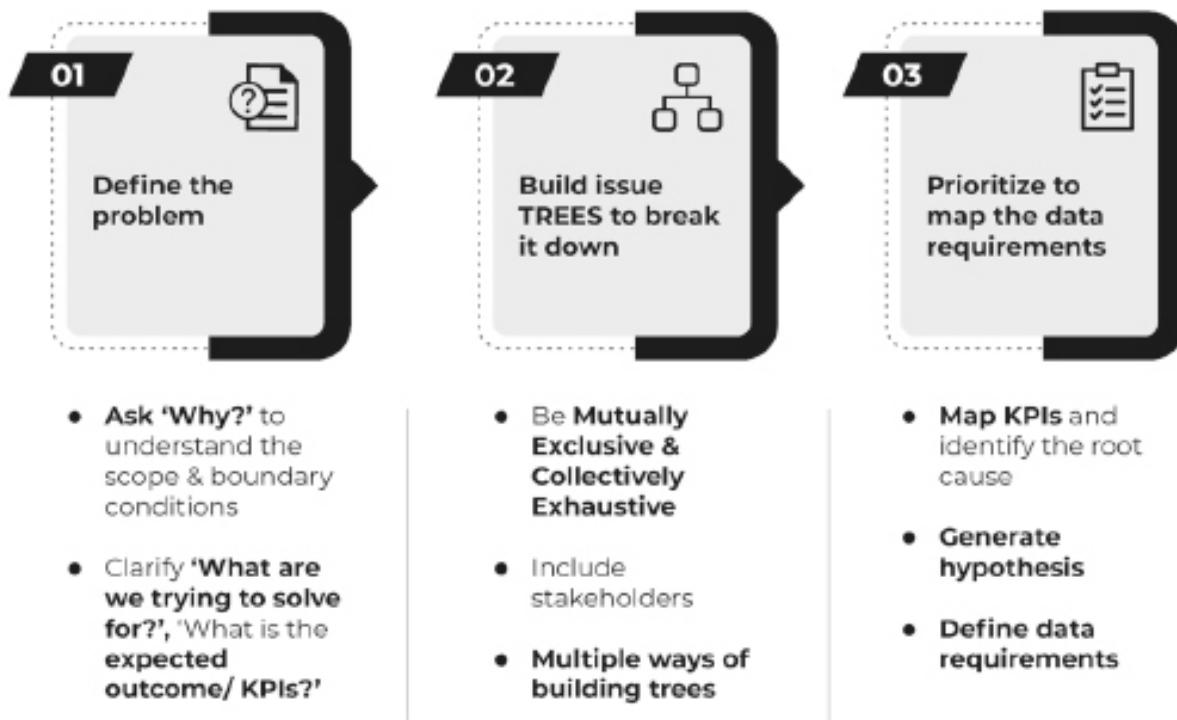
Of course, having a 360-degree view of the customer is critical to having more personal and contextualized interactions with the customers. If done well, it can result in improving customer service, customer trust and loyalty, guide marketing and improve the bottom line. But it needs a focused approach. Otherwise, one could be in a situation of being lost in a raging sea of data with no compass to guide them through. It needs a radical shift in both the mindset and the approach to narrow it down based on the use cases, that will make it as easy as fishing in a pond, vis-à-vis the stormy seas.

For a glimpse of the solution, let's go back to the Flipkart example from the introduction. The data system gridlock we found ourselves in was solved not by our brilliant technologists, but by 'narrowing' down the problem statement—getting to the most critical use cases and thus limiting the data pipelines where we had to ensure data quality.

Three-step approach to problem-solving for data initiatives

How do we avoid problems with data initiatives similar to what we just discussed? I recommend the following three-steps problem-solving approach to get data initiatives off to a good start. Let's explore the process by taking the example of a hypothetical consulting firm XYZ Corp.

Systematic approach to translate business problems into data needs



Step 1: Defining the problem statement

Problems often come to us in broader form, like improving profit margins, optimizing supply-chain footprint, reducing operating costs, etc. These issues are typically the symptoms or manifestations of more complex underlying problems within the organization. The root causes often lie several layers deeper within the organization or its

processes. This means that addressing only the surface-level symptoms may not lead to a lasting solution.

So whenever you embark on the journey of defining the problem, you must start by asking the basic questions first to clearly understand—why? For example, why are we doing this? What is the end goal? What are we trying to solve here? While these look like simple questions, they have the power to refine your thinking and develop a focused and specific approach. So step one is to keep asking ‘why’ until you get to the true reason why it is a problem and why you need to solve it.

At this level you generally do not need too much data. This is the stage where the core team, which includes stakeholders from different functions (business, IT and operations) leading the initiative have to use their industry knowledge, their understanding of the business and their insight into the problem at hand to define the problem more clearly and discuss the outcome KPIs that they are trying to improve. The core team should spend a good amount of time deliberating on the problem at hand. For example, at XYZ Corp., we are asked to improve profit margins. Ask why? Because over the years the profit margin has been consistently above 25 per cent, while in the past six months it has seen a decline of 10 per cent. So, we can now clearly define the problem statement as ‘XYZ Corp. should increase its profit margin by 10 per cent in the next six months’. Why? Because that would help XYZ get back on track. Fair enough. The core team at this point will also discuss the constraints that they might encounter and agree that they must achieve this without compromising on revenue growth. That now becomes a very clear and specific problem statement which has been aligned with the company objectives, understood and agreed upon by all stakeholders.

Here I have purposely used the phrase ‘core team’ again and again because I believe this is not a job for one person or one team. The success of any data initiative hinges on aligning with the key stakeholders in the initial problem-defining exercise. This is because stakeholders from different groups or functions would have their own perspectives on what would be the best way to go about

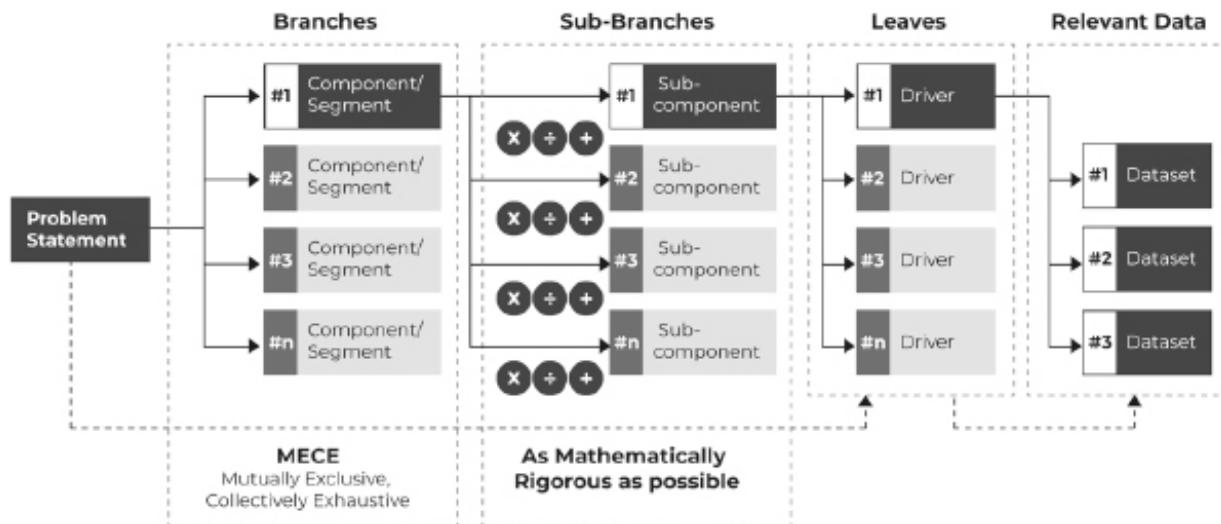
solving the problem. In most organizations, data initiatives are led by a specialized team, which could be data scientists or tech experts or functional specialists. These teams are also responsible for defining the problem statement. This siloed approach can create problems. While the tech and data teams are understandably better placed at driving the infrastructure and analytics aspects respectively, the business and operations teams are better suited to identify the critical business issues as they have the front-seat view to the way the business works.

It is essential to incorporate the cross-functional insights and expertise from the start to reduce the risk of bias and lack of comprehensiveness which can affect the entire data program. So, stakeholders from various teams who will be expected to translate these insights into actions and/or will be affected by the outcome should be brought on board from the beginning.

Step 2: Breaking down the problem into logical components —Trees, trees, trees!

Every problem has some degree of complexity and uncertainty attached to it. Once the problem statement has been clearly defined, we now need to break it down into smaller, actionable components so we can make progress. It is critical for these components to be **mutually exclusive**—each one is separate and distinct, and **collectively exhaustive**—every aspect of the problem is covered under one of these components (MECE).

Building KPI trees is the most effective way of translating business problems into data needs



Taking the XYZ Corp. example further, improving profit margins has two MECE components to it—‘increase revenues’ and ‘reduce costs’. Now what makes up for revenues and costs for a consulting firm like XYZ? The earning from each industry that they serve is the revenue and the cost of delivering the various projects is the cost. But does that give you a clear direction of where to go, or what to do? We obviously can’t improve the revenue of the entire firm, or reduce the costs across all projects now, can we? All it does is provide you with a road sign that indicates where to start digging.

So, these two components must then be further drilled down into the key drivers. The logic tree will at this stage naturally start evolving into a KPI tree. (Or in some cases it might start with broader KPIs like profit margins for various business units that can be broken down into smaller ones, depending on the problem at hand.) The KPI tree, also known as the driver tree, is a structured way of breaking down each component to identify the key drivers that contribute to the performance or functioning of that component. On the far left, you start with the broad problem statement, and as you move to the right (or top to bottom), you continue to break it down, first into key components (branches) and then each component into specific drivers or KPIs (sub-branches). These KPIs

should be actionable, which basically means each one is attached to a specific outcome, and they clearly have a role in solving the larger business problem at hand. Each level of the tree should also be made as MECE as possible. It is therefore important to be very rigorous and comprehensive in this exercise and look at all the possible ways the problem can be broken down.

Additionally, there is no set formula to build a tree. Every problem can be broken down in multiple ways. So, looking at one single tree would not be enough. The core team must do this exercise multiple times building multiple trees. For example, for a service company like XYZ, the revenues can be broken down based on industry, clients, projects or geographies, etc. The key is to be consistent at each level and with each type of tree and not to end up comparing apples with oranges!

This tree-building exercise can easily become too complex. Which brings us to the next step.

Step 3: Prioritize key KPIs to map the data requirements

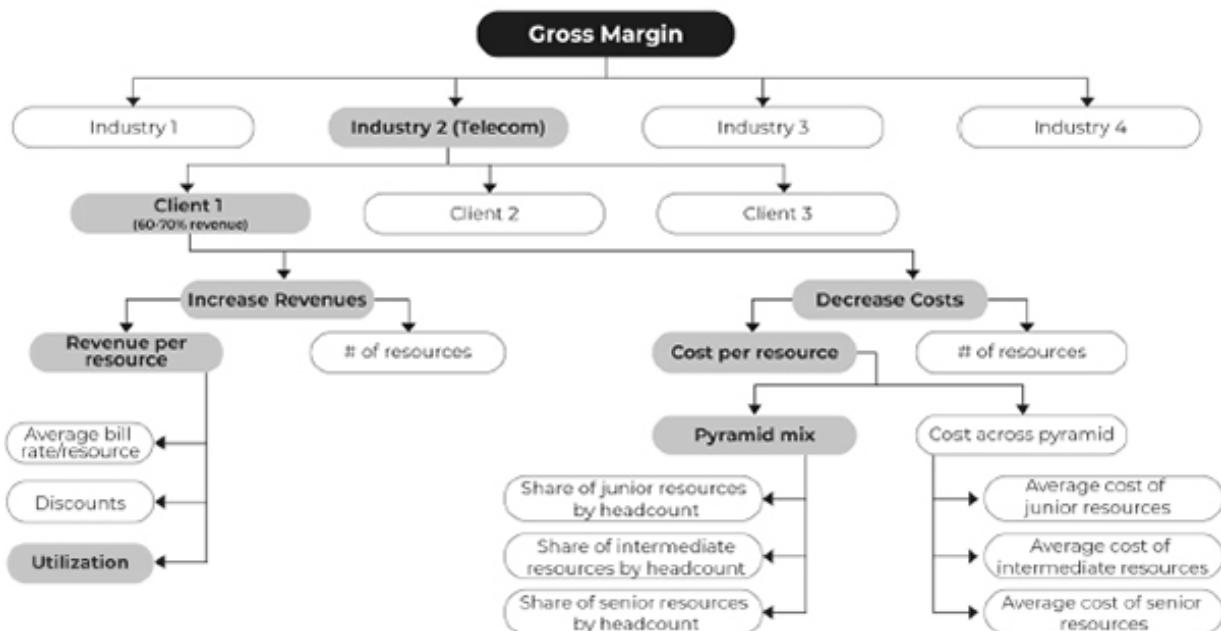
Once you have the specific KPIs and key drivers identified in a MECE fashion, the next step is to prioritize them. For this, you start to benchmark the KPI performance and estimate their impact on the core problem. At the KPI level, you then start to ask questions like 'How important is this KPI to the overall outcome that we want to achieve?' and 'How much can I move the needle on this KPI to achieve the desired outcome?'. Furthermore, you benchmark the KPI performance in the following ways—over time, against internal comparators, against external peers.

With the XYZ Corp. tree, let's say I look at the telecom branch of the tree and look at the client level branch. Here I have a company Client 1 that makes up 60–70 per cent of revenue in that segment and then there is the 'others' category that makes up the rest. So logically I know that any improvement I make in 'others' will not impact the profit margin that much compared to even a small percentage improvement in Client 1's revenues. Therefore, it would make sense to prune the 'others' branch and focus my efforts on the

Client 1 account. In XYZ's case, for example, it emerges that utilization and pyramid mix (see illustration below for reference) is the problem area that needs to be addressed to impact profit margins. I may identify more KPIs in other branches as well that are worth evaluating. The key idea is to get to that level where you can clearly identify the areas of improvement. And while identifying the priority KPIs, it is essential that you use the 80/20 rule—the one or few KPIs that are expected to have the most impact on the problem.

This is where we start defining our data requirements to deeply analyse what is really happening with that particular KPI. Yes, in the earlier steps we do use data to some extent, but mostly it is used as indicators to direct us through the tree-building exercise. But now, at this stage, we have narrowed down the problem to select KPIs that emerge as the root cause, we begin mapping the data required to conduct deeper analysis—build data models to understand why these KPIs are not performing as expected and what needs to be done to bring them back on course.

Illustration - Improving gross margin of a business



Imagine the scope of the broad business problem that we started with versus the specific KPIs that we have now drilled down to.

Doesn't it significantly narrow down the data requirement? While earlier you might be dealing with all the possible data related to profit margins, revenues and costs across board at XYZ Corp., now you are dealing with just a few metrics specific to a few critical KPIs and therefore the data requirements have now become significantly smaller and more specific. That, my friends, is the beauty of building trees.

With the data requirements as well, you continue building trees to further narrow down to specific data elements, depending on the nature of the KPI you need to address. For example, in the case of XYZ, the KPIs—utilization and pyramid mix, for the telecom client 1, you will now map data elements required, like utilization of each resource, share of junior, intermediate and senior resources by each project, average bill rate by service line and so on and analyse these to understand where the scope of improvement lies.

In the Big Data world, where organizations are receiving overwhelming amounts of data every day, this is the only way to systematically narrow down and identify the most relevant ones required to effectively solve any problem, while sidestepping the risk of getting lost or misdirected. It would end up saving organizations from getting into data initiatives that cost too much and take too long, while not achieving the desired outcome.

Key takeaways

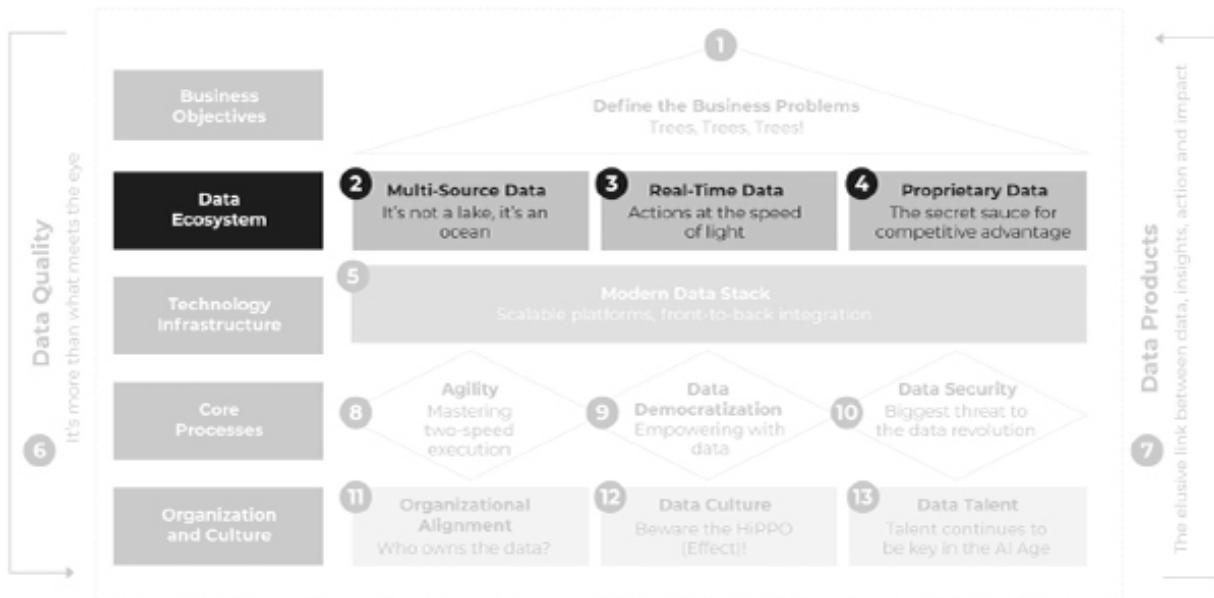
- Problem-solving in the Big Data world requires a fundamental shift from the traditional approach of 'bringing all the data together' to a structured approach of 'narrowing it down to make it more manageable'.
- Clearly defining the business problem and getting to the root cause is the essential first step of the problem-solving process.
- Building a KPI tree is the way to break down a problem into actionable components and prioritize the key issues

to focus on. It also helps to narrow down the data requirements as well.

- KPIs trees should be built as mutually exclusive and collectively exhaustive (MECE) as possible. There are multiple ways of building the KPI tree for a problem, and identifying the most appropriate one is key.

LAYER 2

DATA ECOSYSTEM



'The opportunity presented by the Big Data world'

Multi-Source Data

It's Not a Lake, It's an Ocean

'Variety is the very spice of life, that gives it all its flavour.'

—William Cowper,

Renowned eighteenth-century poet

Variety adds breadth and depth to life. When one is exposed to multiple experiences is when one truly begins to appreciate life and is able put it all together to find deeper meaning and purpose.

Variety provides the multidimensional view critical for learning and growth. In the Big Data world, 'variety' in my opinion, is the most complex dimension of data due to the types, sources and formats that organizations need to manage. But it is the very 'spice' that adds multiple 'flavours' to the insights generated, making it more meaningful and comprehensive.

In this chapter, I will delve deeper into the significance of multiple sources of data and ways to leverage it effectively.

Explosion of sources, especially open data

The explosive surge in data sources in the Big Data world, as discussed in Chapter 1, Data Explosion, is primarily a result of the rapid digitization of modern society. This transformation has been driven by the widespread adoption of smartphones, IoT devices and the ubiquitous presence of social media platforms, all of which generate a vast and diverse array of data, both structured and unstructured.

Moreover, the proliferation of the internet and the growing popularity of the World Wide Web have been instrumental in this data explosion, fuelled further by the open data movement. This movement has led to the exponential growth in openly accessible data sources, like Google's data, public health statistics, census records, weather observations, geospatial information, social media content, data from massive open online courses (MOOCs) and many more.

Which brings me to the larger topic of this chapter—multi-source data. The varied data sources available to organizations can be broadly categorized into internal and external data sources. The internal data sources, generated through the normal course of the organization's functioning, are a critical starting point to build a better understanding of the business. Integrating multiple internal sources and/or combining these with external and open data sources can significantly enhance the quality and depth of insights generated.

Multi-source data: The value multiplier

In an organization, multiple functions come together and work towards achieving common goals or objectives. Organizations also operate as a part of a bigger ecosystem that consists of suppliers, sellers, channel partners, regulators, customers and more. All these entities, both internal and external, affect an organization's day-to-day operations and performance. In addition, political, economic and environmental factors also have the potential to affect the organization. And as organizations today operate in the VUCA world, the impact of these factors has further intensified. It is therefore critical for organizations to take both internal and external factors into account to build a comprehensive view for decision-making.

Going beyond the data silos to integrate multiple internal data sources can help organizations build a better view of their business. Furthermore, combining internal data sources with multiple external data sources can help organizations generate deeper insights by providing a more comprehensive view of competition, industry

trends and their customers. Leveraging multiple data sources also enables organizations to validate their findings and diminish the effects of biases. Integrating multiple external and internal sources enables organizations to achieve multiple goals:

Connecting the dots: Integrating multiple sources of data enables organizations to fill in the information gap and provide a multidimensional view of its operations. It helps connect the dots to make better decisions, improve efficiency, reduce risk and drive growth. A company like FedEx is known for leveraging external weather data and integrating it with its internal logistics systems to reroute shipments when adverse weather conditions are detected.

Understanding the customer better: With the help of multiple sources, companies can gain a deeper understanding of their customers, their preferences and their needs to provide better customized products and offerings. DBS, one of the premiere banks of Singapore, is leveraging multiple data sources, both internal and external, to provide hyper-personalized offerings to its customers to help them make more informed financial choices.

Addressing the bigger picture: Today, organizations also bear the responsibility to contribute to the betterment of society and the planet. Leveraging multiple external and internal sources is critical for organizations to be the agents of change. Unilever has always been a leader, working for years towards building a better tomorrow. Their 'data for good' team is combining social media and internal surveys and leveraging external reports and case studies to identify current and potential challenges faced by employees with disabilities at Unilever.

In the Big Data world, there is a lot more data generated today, most of which is available through multiple sources outside the organization. If leveraged the right way organizations can generate newer, deeper and comprehensive insights which would not be possible by relying on just one or a few siloed data sources.

Organizations are rapidly realizing this, as nearly half of companies reported using external data in their analytics activities, according to a recent survey. Having said that, integrating internal data sources itself is not an easy task as more than 80 per cent of the organization still struggle due to multiple data silos that exist across the organization. Additionally, a 2019 survey indicates that about 92 per cent of data analytics professionals feel their companies need to increase their use of external data sources.¹

Multi-level integration of data

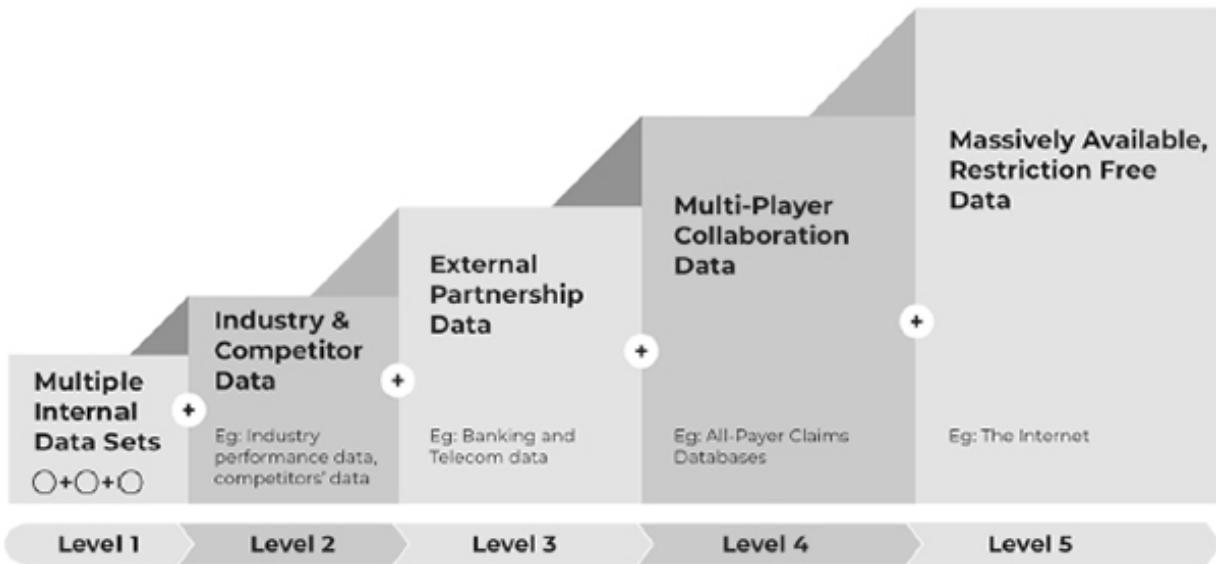
Leveraging multi-source data is like progressing through the various levels in a video game. Each new level adds to the complexity, which requires higher skills and capabilities but results in greater gains and rewards.

Let us look at data at each level and understand the value it adds:

Level 1: Internal data

The immediately available data with any organizations are the internal data sources. It is data generated internally, from across the value chain of the organization, like corporate data which includes sales, profitability, etc.; data from operations like customer interactions, production, efficiency, etc; human resource data like employee productivity, employee satisfaction, hiring and attrition and so on. It is the primary source of information on how the organization is performing, and how various functions across the value chain are contributing to that performance. Collecting and analysing internal data helps an organization benchmark its own performance against previous years as well as assess its internal strengths, weaknesses, threats and opportunities on a regular basis. It enables companies to create actionable plans to capitalize opportunities and mitigate risks.

There are multiple levels at which organizations leverage multi-source data



In my roles at Fidelity and Flipkart, I was responsible for both strategy and analytics. The biggest gap that I recognized at both these companies, which is still predominantly the case across companies, was a disconnect between the strategy and analytics processes. Analytics teams at these organizations were building data-driven, granular-level customer segmentation models, while most times the strategy team would formulate strategies for growth, market entry or portfolio optimization, etc., largely based on high-level customer segmentation, which is more market-feedback driven and not as granular or data-driven as what the analytics team has developed. This is an obvious missed opportunity. Integrating operating data, customer data and other internally generated data into the corporate strategy decision-making can add significant value with deeper and much more granular insights.

Another example that highlights the significance of combining multiple internal sources is how customer decisions are made in multi-product companies, which we see in many of our banking clients. Typically, customer data from various product groups like personal loan, auto loan, credit card, etc., are managed in silos. As a result, a customer defaulting on a credit card could be getting offered a personal loan or auto loan because of the lack of

integrated view. This gap can be addressed by integrating data from various product groups to make better decisions on product offerings and reduce risk.

Level 2: Industry and competitor data

While combining multiple internal data sources is essential to get a comprehensive picture of an organization's internal value chain, not many companies have realized the full potential of combining it with the external market data or competitor data yet. It is essential to put a company's performance into perspective in comparison to the industry it operates in and its competitors, especially so in a world where organizations are a part of a complex ecosystem.

Industry data includes aggregated industry-wide information, which includes industry growth, profitability, sub-segments, regions, competitors' positioning, market share and differentiation, etc. It also includes new regulations, policies and industry-specific changes. Evaluating against industry and competitor performance helps companies benchmark their performance against the industry or the competitors and set more realistic goals and targets. Identifying key trends that are shaping the industry can help build a more growth-aligned and profitable product portfolio and enhanced offerings. It also helps organizations recognize the potential challenges and threats which can be proactively mitigated. Additionally, the data on operations and performance of the major competitors in the market helps organizations promptly identify the gaps in their offerings. It helps formulate strategies to deal with new competitive developments and identify potential risks in time.

When I joined Fidelity International, one of the largest investment management firms in the world, it was going through a business transformation. The firm had experienced a significant decline in sales from between 2010 and 2012, and growth was elusive. Upon investigation I found that at Fidelity the sales performance was being measured predominantly based on internal data—comparing movement of funds using historical sales data—with limited competitor benchmarking. While this methodology was fine for the

firm to track its sales movement through time, it wasn't comprehensive enough to build segment-level strategies required to win in a very competitive market. So, we introduced a new three-dimensional evaluation criterion to infuse the outside view. First, performance or growth of the segment we are operating in; to understand the market growth in granularity. Second, market share growth of each of our products; to understand how well our products were faring competitively. And third, investment performance; to evaluate our products' performance, vis-à-vis other similar product. Equipped with better understanding of our product portfolio against the market opportunities and our competitors, we were able to better identify the right segments to focus on. This became the basis of revamping our strategies for our product portfolio and which resulted in higher growth over the next few years.

Level 3: External partnerships data

Organizations do not operate in silos, rather they are part of a larger ecosystem, which includes suppliers, clients, vendors and players from adjacent markets—products or services that are situated upstream or downstream. These players also significantly impact the organization's performance. Incorporating data from such businesses enables the organization to reach beyond their walls to understand how customers are interacting with their brands. This helps in identifying opportunities to enhance the products and offerings and improve customer experiences. For example, a consumer-packaged goods company may have data on their own customers and their purchase history but will not have access to their overall buying behaviour or purchase patterns in terms of other goods and services. Similarly, data from the adjacent markets can enable companies to expand their offerings in terms of new markets, new products, new customer segments or new channels.

Mobile wallets is one of the most innovative partnerships between banks and telecom operators that has revolutionized the way people around the world transact today. By the year 2020, Samsung had

launched Samsung Pay in twenty-six countries. But it hit a roadblock in Germany. The German banking market was highly fragmented, with too many banks to deal with and integrate with Samsung Pay. Also, the legacy model of relying on credit cards did not work in that market as credit card penetration was comparatively low. Therefore, Samsung collaborated with Visa and engaged Solarisbank, Europe's leading Banking-as-a-Service platform, to develop a virtual debit card with Samsung's branding, eliminating the need to set up individual contractual relationships with dozens of German banks. Samsung also integrated Solaris' innovative identification method called Bank Ident, enabling the identification of Samsung's users through a TAN-approved (using a one-time password) microtransaction from their existing bank account and confirming their registration with a digital signature. This enables Samsung to onboard thousands of customers simultaneously, 24x7.² Other examples of such partnerships are the hotel-industry players collaborating with the airline industry to provide loyalty rewards, and cross-sell vacation packages, or data-sharing between banks or financial institutions to design better product offerings and lower default risks.

Level 4: Multi-player collaboration data

Multiplayer collaboration data is a pooled data set that is accessible only to those members that have agreed to collaborate, leveraged to jointly solve a business problem. For example, original equipment manufacturers (OEMs) share supply chain data to move, track and manage container and part inventory between companies more efficiently.

When the business goal or impact is much greater and cannot be achieved by one organization alone, it collaborates with other players in the markets to pool resources, especially their data pool, to build capabilities large enough to tackle the business problem effectively. This type of partnership not only benefits all the participating entities but also helps make significant advances in the

industry, or to achieve a public good. Today, most industries have formed industry consortiums—where representatives from several different companies of the industry come together to pool resources and knowledge to achieve common objectives.

The all-payer claims databases (APCDs) is a great example of such an initiative. We are all aware of the rising healthcare costs in the US and the ongoing battle to curb these costs, which is a hot issue in the US today. Considering the highly complex and fragmented US market, it is not feasible for any one organization to control the rising healthcare costs or tackle healthcare disparities. Therefore, for a more comprehensive picture of the healthcare delivered to the residents eight states have created, and thirteen states are in the process of creating APCDs. This database collects and aggregates data on payments made by commercial health insurers, by self-insured employee benefit plans, and the Medicaid and Medicare programmes. It is beneficial to multiple stakeholders, including policymakers, consumers, payers and researchers. It helps increase healthcare spending transparency and enable informed decision-making. Though this initiative is new, the APC council has already reported more than forty research studies on the impact of APCDs, where multiple stakeholders have reported benefits like more informed rate-review process, premium cost evaluation, identifying preventive and routine healthcare followed by the healthy population, etc.³ Similarly, Global Credit Data (GCD) is a non-profit global-data consortium formed by fifty-plus banks that collects anonymized internal data of member banks to help them build credit-risk models to manage their risk exposure better. Or the Independent Data Consortium for Aviation (IDCA), which has leading aviation companies at all levels of the industry coming together to identify and develop guidelines, rules and standards on data sharing in non-competitive manner among stakeholder groups.

Level 5: Massively available restriction-free internet data

Organizations today also have access to data which is open for anyone and everyone to access, modify, reuse and share. For example, government data or data from international institutions like the World Bank, the US Bureau of Labor Statistics, OECD (Organisation for Economic Co-operation and Development), IMF (International Monetary Fund) and others. In addition to that, the tech-savvy customers today find multiple ways to interact with the virtual world. They are also generating huge amounts of content online. One such emerging source is social media, where information is available freely and in abundance. It is therefore rapidly becoming the key source for gathering consumer insights. Another good example of openly available data, is the massive open online courses (MOOCs) which can be leveraged by organizations to build better and deeper business insights or improve data and analytical models by leveraging online courses on topics like machine learning, AI, programming, etc.

About 99 per cent of the Fortune 500 companies are actively present on social media.⁴ They are leveraging social media to market their brands, interact with their customers, understand their target market and improve their product offerings. One such example of companies leveraging such freely available data to generate insights and translating these into top-notch services and offerings is Netflix. Through social-media listening, it discovered that one of the issues highlighted by viewers was that they would fall asleep during a binge-watching session and the videos would continue playing throughout the night. To tackle this issue, Netflix introduced Netflix Socks—a built-in sleep-detection system. It would start flashing a red light in case viewers were starting to fall asleep to let them know that the show would pause.

Amazing is the extent to which data has enabled organizations to provide innovative solutions to delight their customers!

Integrating multiple data sources: A colossal task

Over 80 per cent of enterprise business operations leaders say data integration is critical to ongoing operations. Also, 67 per cent of enterprises are relying on data integration to support analytics and BI platforms today, and 24 per cent are planning to in the next twelve months.⁵

According to another survey, on average, an organization collects data from 400–1000 data sources to feed into their analytics systems.⁶ This number is expected to grow rapidly as the number of both internal and external sources continues to compound, making the process of integrating them even more daunting. Few common challenges that come up when organizations are trying to integrate multiple sources are:

Data heterogeneity challenge: It is the most obvious and the biggest challenge with multi-source data. Multiple data types have high variability in terms of data types, formats and frequency. This adds to the complexity and time consumed in pulling all the data together and bringing it into a compatible format for integration. For example, you might need to integrate data from web API, CRM databases and e-commerce websites which would all be in different formats.

Data integrity issues: Data when sourced from multiple sources may result in duplicates or even worse, the data from two different sources may be at conflict with each other. Furthermore, chances of erroneous data, missing data, or incompatible data becomes greater and multiple sources are brought together for processing. This poor quality of data affects the integration process as it requires more time to clean and validate data.

Lack of scalability: As the business problem grows, and the number and types of data sources increase, the complexity of getting data from multiple sources intensifies. Building a reusable,

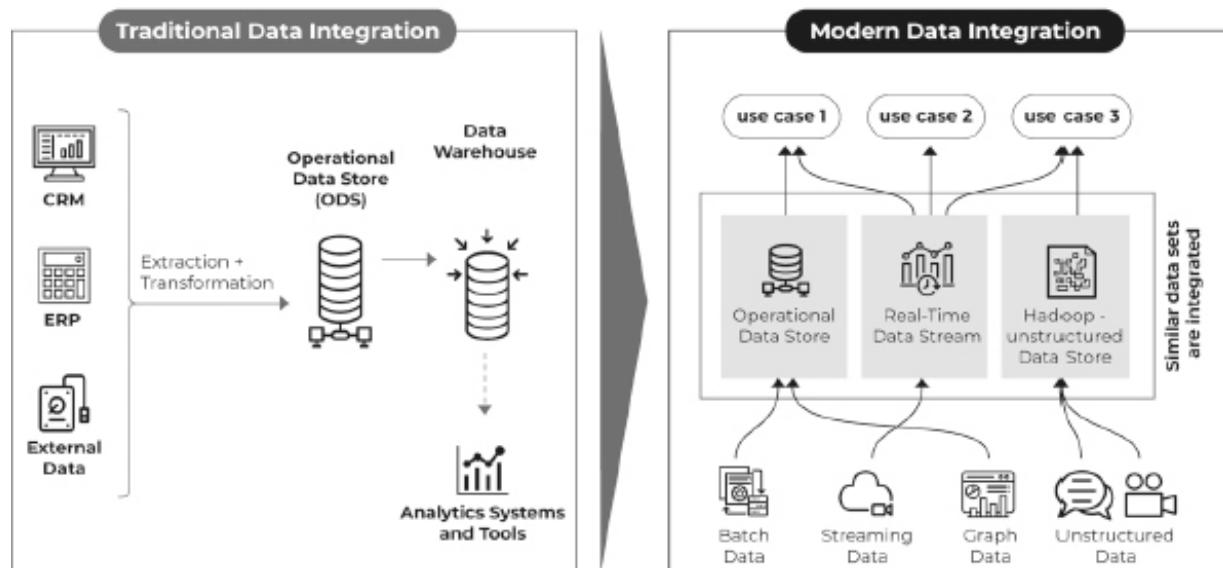
scalable data integration system taking current and future possibilities into account is a challenge.

Time-consuming process: As the volume, variety and velocity of data continues to explode, the efforts of bringing all this data together with the view of integrating it all to generate business insights is a humongous task. The merging process can take too long and can become a tedious task if not done with proper planning and realistic expectations from the beginning.

More 'logical' than physical approach needed

The traditional approach of an organization towards data integration from multiple sources has been to build a staging layer of operational data store (ODS), where all kinds of data would be integrated together at one place. But the scope of data, in terms of the 3Vs—volume, variety and velocity—continues to grow at a blinding rate. So, the complexity around integrating all of it in one place in a traditional ODS has heightened, as integrating such very different types of data is very difficult.

Traditional data integration anchored on 'Bringing it all together' to modern data integration approach 'Logically as per User Case'



Organizations need a different approach to make this process more effective and manageable. The solution lies in breaking down the unlimited number of data pipelines based on the type and nature of data and logically connecting them based on use cases that these are expected to feed. A concept that I have delved into in detail in the beginning of this section in Chapter 7, Define the Business Problems.

Multiple different data stores need to be created to make the integration process more efficient and effective. For instance, real-time data that requires a specific type of processing done in micro batches is stored separately from batch data, where processing can be done in bigger batches. So the data stores are strategically designed based on the type of data and business requirements. This data architecture approach of building distributed data stores depending on the business unit or the type of data is known as 'decentralized data architecture'. And that is where the recently growing popularity of 'data fabric' and 'data mesh' type of architecture has come into play.

Data fabric is an architecture where data is stored in different data stores based on type and managed through a strong data cataloguing and governance layer. Think of it like a large tapestry made up of different threads of different colours. Each thread represents a piece of data, which can be weaved together on a loom with other threads to create beautiful patterns as per our choice. In such a distributed architecture of data fabric, first-level integration happens within the datastore for a specific type of data set. Integration of data from these datastores then happens based on the use case as and when needed.

Data mesh, another popular approach, is an architecture where data is stored based on business units, with teams having full access and control over their data and data products. It's like a symphony orchestra, divided into different sections. Each section plays specific parts of the composition, but each must be played in harmony with other sections to create beautiful music. The key point is that owing to the explosive growth of data, organizations have to approach the

data-integration process more logically than trying to physically bring it all together, which would make it a mammoth task.

Key takeaways

- With the continuous emergence of diverse and novel data sources rapidly adding to the 'variety' dimension of Big Data explosion, organizations have the opportunity to leverage abundant data sources to generate deeper, more meaningful and comprehensive insights.
- In a world where organizations operate as a part of a wider ecosystem, integrating various internal sources and combining these with external data, at multiple levels, can provide organizations a multi-dimensional view essential to enhance their operations, improve customer experience and drive greater impact.
- Since integrating multiple data sources can be a daunting task owing to the heterogeneity of data, organizations must move away from the traditional approach of 'bringing it all together' to a more logical integration approach based on use case—an approach that is gaining popularity with data mesh and data fabric type of architecture.

Real-Time Data

Action at the Speed of Light

'A new beauty has been added to the splendour of the world—the beauty of speed.'

*—Filippo Tommaso Marinetti,
Italian poet, editor, art theorist and founder of the Futurist movement*

In the earlier chapter on multi-source data, I focused on how organizations can leverage the 'variety' dimension of data to their benefit. In this chapter, let's address another dimension of data —'velocity'—the speed at which data is generated today.

The world operates in real time today!

Throughout history, people devised various methods for real-time communication. Ancient Chinese used smoke signals on the Great Wall, Native Americans conveyed messages with smoke, which signals the selection of a new Pope even today. In the eighteenth century, Paul Revere's lanterns warned of British arrival. In Africa, 'talking drums' mimicked speech rhythm. Then in the 1800s, following the invention of the postal and telegraph system, dots and dashes of the Morse code were used to deliver messages across the electrical telegraph system. In 1876 the meaning of real time changed forever with the words, 'Watson, come here, I want to see you', uttered by Graham Bell over the telephone. And then came computers, wireless cellular networks and the internet, transforming real-time communication from 'human-speed' to 'machine-speed'.

Today, thanks to the evolution of technology, we live in the world of instant messenger, live streaming and ten-minute deliveries. We are connected to the internet 24x7 one way or another, be it via our smartphones or through smart devices. Alexa and Hey Google have now become household names while Fitbit, Apple watches and other fitness trackers have become a lifestyle choice for many. Real-time information is no longer generated in special circumstances. It is omnipresent!

An average person today creates 1.7 MB of data every second, through digital interactions like online activities, social media usage and others.¹ Which means 17,82,579 bytes of data per second (1 MB = 10,48,576 bytes). Now multiply this with approximately 5.3 billion users on the internet globally.² So, the estimated amount of data created globally every second (~9,447,668,700,000,000 bytes) is about 9.2 quadrillion bytes per second. Isn't that mind-boggling?

Adding to that, according to IDC estimates, there will be 55.9 billion connected devices worldwide by 2025.³ These would include media devices like smartphones, laptops, wearables, other applications like smart grids, autonomous vehicles, asset monitoring and tracking, etc. IDC further estimates that 79.4 ZB of data will be generated from these connected devices by 2025.⁴ Imagine the velocity at which data will be generated from all these connected devices.

What is real-time data?

Real-time data can be defined as 'any data that is processed as soon as it is created and captured and is converted into actionable insights rapidly'. In other words, it is the data that is available as soon as it's created and acquired. And instead of being stored for later use, it is redirected to the users immediately, critical for instant, live decision-making or for driving automated actions. It's like rapidly flowing water, in contrast to the traditional data which is similar to a water reservoir.

Real-time data is captured in two forms:

1. **Streaming data:** This data has no definite beginning or end, hence streaming like a river. The data is generated continuously from thousands of sources, which is sent in small batches to the data records simultaneously. For example, log files generated from mobile or web applications by customers, gaming activities, social media feeds, geospatial services data, etc. This data needs to be processed continuously and sequentially on a record-by-record basis. It can be used for a wide range of analytics like correlations, aggregations, filtering and sampling. This information can be leveraged by companies for various purposes like understanding the customer usage and activity patterns, and continuously track business activity, to be able to promptly respond to any developing situation.⁵
2. **Micro-batching:** Micro-batching involves processing data that has a clear start and end point, with minimal batch size (often just one). This approach speeds up data management by loading data frequently, even in small increments like seconds. These micro-batches can be processed at high speed, approximating the real-time data at the source. It is very similar to traditional data processing, except this happens in very small batches that are processed within seconds. Micro-batches enable near real-time use cases that do not require to be up-to-the-moment accurate. For example, IT-systems monitoring, or processing sensor data may not be real-time but may be updated every few minutes, as per requirement.

The fundamental shift

In today's world, where the speed of action is paramount for business, organizations do not have the luxury of time to collect all

the data and wait to make decisions based on insights generated over time. According to a McKinsey survey, high-performing businesses are 5X more likely to use real-time data compared to counterparts.⁶

Through the years real-time data has been used across multiple industries where the consequence of non-response or delayed response is high both in terms of risk and cost. For example, financial institutions have been using such data to track fraudulent transactions and stop them even before they happen, resulting in millions saved in tracing, cancelling and rectifying damages caused. The airline companies are highly dependent on real-time data for operations, from deciding when to close the flight doors to delivering baggage status to their customers. In recent years, the role of real-time data has expanded significantly, to creating experiential moments for the customers like virtual shopping assistants that act as store associates in real time; or building newer business models, like ride-sharing apps, food-delivery apps, etc. Companies like Facebook, Uber, Airbnb, Amazon and Google maps, offer just a few of the many services, hinging on real-time data that consumers increasingly rely on throughout their typical day.

As a result, the data landscape has seen a fundamental shift in recent years as more and more data is generated, processed and consumed in real time. And due to the instantaneous nature of real-time data, the data-management cycles have significantly compressed. Traditionally, data-processing was done in batches involving collecting, storing, cleaning and analysing data at a later stage, with results available only after the data-management cycle was complete. However, real-time data is processed and consumed immediately.

Real-time data and compressed data-management cycles has led to growth in real-time analytics that processes incoming data instantly, allowing users to analyse and comprehend it as it arrives. This approach contrasts with traditional analytics that rely on historical data for future predictions. Real-time tools swiftly handle large data volumes, offering rapid insights within seconds or minutes

of data entry. This instant feedback supports on-the-fly decision-making and automation, addressing current business needs effectively.

Today many data-driven applications and solutions based on real-time analytics, have been implemented by organizations, for themselves and their customers. Whether it is the personalized recommendation based on browsing activity on an e-commerce website, or fraud detection in real time by financial institutions, or monitoring patient vitals in real time in healthcare, all this has been made possible only through the implementation of real-time analytics using advanced tools and technologies like distributed computing systems, in-memory computing, edge computing, etc.

How are leading companies leveraging real-time data?

Real-time data enables organizations to keep a timely check on how their businesses are performing and how external factors are affecting their operations in real time. Organizations can leverage this information to drive superior outcomes across their value chain. The cases where organizations need to instantly translate data into insights, and insights into actions is where real-time data is critical.

Some of the areas where organizations have been leveraging real-time data are:

- 1. Personalizing customer experience:** Real-time personalization delivers customized content or service to each individual user, while they are interacting with the brand. This can be done through instant messages, mobile apps, websites or other marketing channels. Coupling the right data with real-time information is critical to create that superior customer experience. An obvious example here is Amazon. Its recommendation engine recommends products from different categories based on what the customer is browsing now and pulls up those products in front of you and helps you draw comparisons based on price, reviews, rating and features

of that product with closely matching products in that category. It enables Amazon to increase the average order value by upselling and cross-selling products through this process. About 35 per cent of [Amazon.com](#)'s revenue is generated by its recommendation engine.⁷

2. **Improve customer service:** In the age of ten-minute delivery, customers do not want to wait forever to get their issues resolved or their queries answered. A survey indicates that about 75 per cent and 78 per cent of Facebook and X (formerly Twitter) users, respectively, expect a response to their queries within an hour.⁸ And another indicates that 68 per cent of consumers say they are willing to pay more for products and services from a brand known to offer good customer service experiences.⁹ Use of chatbots and live chats by digital banks is a good example here. These digital-only banks complement their live chat facility with chatbots to assist users with basic financial tasks like managing their accounts or making transfers and getting answers to the most frequent questions, any time, anywhere. China's WeBank is one such successful example where it has extensively deployed AI in user interactions and claims that almost 98 per cent of customer queries are addressed by AI chatbots.¹⁰

3. **High-velocity decision-making:** Gathering and processing data quickly, armed with artificial intelligence, machine-learning algorithms and automation, organizations are able to assess options and respond faster to rapidly changing market and customer demands. Having real-time visibility into all aspects of the business has become increasingly crucial for organizations to take informed decisions at a faster pace. The enormous potential of real-time data not only gives businesses the agility they need in decision-making but also provides proactive insights into their operations,

potential risks and opportunities for improvements. For example, AWS rolled out a supply chain management application in November 2022, AWS supply chain, that provides enterprises with a unified real-time view of their suppliers, logistics, inventory and more. Powered by ML, it helps enterprises unify and analyse data from across multiple ERPs, supply-chain systems and vendors in real time, enabling them to maintain visibility at all times, and quickly respond to potential supply chain disruptions. The aggregated view of data, automatically analysed and presented through visual, interactive dashboards enables faster decision-making and quicker actions towards building a resilient supply chain.^{[11](#)}

4. Real-time monitoring and rectification of issues:

There are certain high-risk use cases where it is critical to monitor the processes on a continuous basis to avoid disruption or huge losses. These use cases require most current up-to-date data to be captured on a continuous basis. A prominent use case is the monitoring and maintenance of industrial equipment in real time, using IoT sensors, which enables predictive maintenance. High-risk operations like oil and gas rigs require round-the-clock monitoring of numerous equipment like pipes, valves, wellheads, tanks, etc., and parameters like temperature, vibration, pressure, flow rates, corrosion, gas leaks, etc., which is not possible to do manually. For example, Exxon Mobil, an American multinational oil and gas corporation, has launched Mobil Serv Real Time. It is an oil-condition monitoring tool that enables instant, remote access to detailed oil diagnostics. It provides a real-time data stream on oil health and alerts as soon as issues are detected. Updates and analysis are accessible through a live dashboard. It facilitates continuous monitoring of oil health against multiple parameters, prevents equipment failures and maximizes asset

utilization.¹² Other prevalent examples are the use of real-time data for fraud detection by financial institutions and social-media companies to identify and remove questionable content from their websites.

5. Be agile to respond to external changes quickly:

Real-time data enables organizations to quickly adapt to external disruptions. And if recent history has taught us anything, it is the importance of utilizing the data at the right time—in real time. During the pandemic, the majority of organizations were struggling to make sense of how events were transpiring in real time around the world. During this time of chaos and uncertainty, many organizations had to rely on real-time data to either survive or to save lives. Janssen Pharmaceuticals (Johnson & Johnson) is one such organization that undertook an enormous data-science effort to help guide its potential vaccine research. It used data in real time to track the pandemic and determine the next hotspots. Janssen built a global surveillance dashboard that pulled in data at a country, state and even county level. This helped it determine the next location where it should test its investigational Covid-19 vaccine candidate. Tracking the spread helped Janssen understand the travelling pattern of the disease to strategize better.¹³

So, where does the challenge lie?

The burgeoning ‘velocity’ at which data is being created is becoming increasingly difficult for organizations to manage. Therefore, not many companies are able to leverage real-time data to its full potential. The issue is compounded especially when some of this data comes with an expiry date. If not analysed and used immediately, such data may lose its validity or importance and organizations, as a result, may miss out on critical insights essential for quicker decisions or delivering superior customer experiences.

According to the Compliance, Governance and Oversight Council (CGOC), a community of more than 3800 leaders and practitioners from legal, IT, information management, privacy and security, 69 per cent of all corporate data that is collected has lost some or all of its business value,^{[14](#)} as organizations fail to identify its use in time. It is therefore important for organizations to utilize data quickly and efficiently while it is still valuable. Walmart—the world's largest retailer—knows this very well. It has built the world's biggest private data cloud, capable of pulling in 2.5 petabytes every hour. But that data only represents transactional sales made over the last few weeks. If Walmart is unable to get insights from this data while it is available, it results in lost opportunity for it.^{[15](#)}

So, it becomes a double whammy for organizations where they not only have to keep pace with the high velocity data, but also have to find ways to utilize it quickly before it loses value.

New and emerging technologies 'to the rescue'!

In a world where 328.77 million terabytes of data are created every day,^{[16](#)} storing and analysing such high-velocity data on a continuous basis is a mammoth task, where traditional data management approaches, designed to deal with batch data, are ineffective. Organizations need the right technology infrastructure to support such a high velocity data management value chain. This has been made possible by the newer and more efficient technologies that are available today. Here are the challenges at each stage of the data-management cycle and the technology solutions that can help organizations become more effective with real-time data.

Stage 1: Data ingestion

The large volumes of data generated continuously at a blinding rate can overwhelm traditional batch-processing-based data systems, which are not designed to work with time-sensitive and continuous flow of data. In such a scenario, solutions like *Amazon Kinesis*^{*} and

Kafka are more useful. These are designed to ingest data streams continuously by transforming them into smaller pieces of data records that are arranged sequentially by arrival time and distributed across systems. As the volume of data goes up, the number of pieces also go up, providing the required scalability and speed.

Stage 2: Data processing

Processing real-time data is resource-intensive, as it requires immediate and continuous processing, which becomes difficult through traditional data management approaches that lack speed and scalability. Tools such as Apache Spark^t that are built on a distributed computing framework, where tasks are divided and distributed across multiple computing nodes (systems) for faster processing, can help in processing real-time data.

Stage 3: Data storage

The storage technology also needs to evolve to facilitate the nature of real-time data. Traditional technologies like relational databases—databases that store and provide access to data points that are related to one another, do not work with real-time data because speed and scalability can become a constraint. To address this, cloud-hosted NoSQL real-time database solutions like MongoDB are now available. These are ultra-low latency, flexible and scalable platforms for real-time data storage. NoSQL databases can be horizontally scaled, which means that they can handle large amounts of data by adding more servers to a cluster and NoSQL databases have flexible data models making them suitable for processing data that can be highly diverse and constantly changing.

Stage 4: Insights generation

Insight generation on real-time data is instant and continuous in nature. In a traditional approach, typically batch data is brought together in a central warehouse and insights are generated on that,

making it a time-taking process. Tools based on ‘in-database analytics’ technology allows querying and processing of real-time data to be conducted within the database by incorporating analytical logics in the database itself. And accelerated insights generation approaches like feature stores^{*} are designed with predefined data features for identifying and using the data needed for running ML models from real-time data streams.

Stage 5: Real-time action

Organizations should be able to leverage real-time data to drive action in real time or near real-time. But traditional data architecture is designed with focus more on insight generation rather than driving immediate action. Hence driving action in real time is not made possible. Technologies like edge computing help solve this by storing, processing and analysing data closer to where it is generated to enable rapid, near real-time analysis and instantaneous action.

The rapid growth and availability of these technologies has enabled organizations to leverage real-time data to make critical business decisions and take actions in real or near real-time.

But is real-time data good for every use case?

It is true that organizations are increasingly realizing the benefits of using real-time data to make faster decisions and create superior customer experiences. Many of them are seeing positive results as well. According to a 2022 survey, 80 per cent of organizations surveyed reported revenue uplift due to real-time data analytics and 62 per cent of companies reported more efficient process rollouts after implementing real-time data systems, which indicates that companies have been able to leverage real-time data to improve performance.¹⁷

But at the same time, another survey revealed that 87 per cent of organizations fell short of their budget for analytics engagements

and real-time data use cases cost significantly more than batch data use cases for development. [18](#)

So, while there is abundance of real-time data everywhere, and there are newer, more efficient technologies available to capture and process it, it is neither feasible nor advisable to use such data for every case. This is because, as I discussed earlier in the chapter, capturing and storing real-time data is a humongous task that requires very high investment and high-end technologies.

So, how do you identify the right use cases?

To narrow down the use cases most suitable for real-time implementation, organizations must focus on those where the insights generated can be utilized to create 'automated actions'. That is key to realizing the full potential from real-time data. They can further identify the right use cases by evaluating the trade-off between 'impact' and 'effort'. Organizations must prioritize use cases that are very high or high on impact where the efforts required to capture and process real-time data are justified by the impact achieved. So, what do I mean by impact and effort?

Impact: The value of real time is realized when real-time decisions are enabled, and immediate actions are taken based on these decisions. There are three types of impact that would justify the use of real-time data. One, when the real-time data creates significant value for the end customer and helps deliver superior customer experience. A classic example is real-time personalized product recommendations. Two, when a business use case requires monitoring operations in real time to improve the efficiency of the business. For example, real-time monitoring of equipment to predict repair and maintenance requirements. And three, use cases where speed of action is paramount to mitigate risk. For example, proactive blocking of a potential fraudulent credit card transaction as soon as it is triggered. The impact of leveraging real-time data has been discussed in detail in the segment above as well.

Effort: The other aspect that organizations must consider while determining real-time use cases is the 'effort'. Organizations must consider the investments, both in terms of money and time, that they will have to put in to build a real-time use case. Two types of cost should be evaluated against the efforts to identify the use cases. One, infrastructure cost. Compared to batch data, the cost of building capabilities to process real-time data requires huge upfront investments and continuous investments to run and maintain such systems. So, the opportunity cost associated with investing in the infrastructure to leverage real-time data—forgoing other potential opportunities or benefits that could have been obtained by investing elsewhere—is very high and organizations must be mindful of the trade-off that they are making to maintain it. Two, while considering investing in such technologies, organizations must evaluate the extent of changes and the extent of integration required to their current technology ecosystem. For effective technology integration, it is critical that the current technology ecosystem be upgraded. Unless that happens, organizations will fail to achieve the desired impact and would end up making unnecessary investments, adding to the already high cost of implementation.

So careful evaluation of the trade-off between impact and effort would help organizations identify the most critical use cases that require leveraging real-time data.

Key takeaways

- Real-time data is exploding and if used well can be a game changer for organizations—improving quality of decisions, improving customer experience, managing risks, offering superior agility and much more.
- Real-time data has brought about a fundamental shift in the way data is processed. It has made the entire process instantaneous, collapsing the data-management cycle.

- Organizations must identify and implement the right technology infrastructure throughout the data-value chain, capable of managing the dynamic and instantaneous nature of real-time data.
- Owing to the sheer volume and velocity at which real-time data is generated, it is neither feasible nor advisable to use real-time data for every use case. Organizations must carefully evaluate the effort vs impact trade-off to prioritize high-impact use cases.

Proprietary Data

The Secret Sauce for Competitive Advantage

'In order to be irreplaceable, one must always be different.'

*—Coco Chanel,
French fashion designer and founder of the Chanel brand*

The signature style

Leonardo da Vinci, Michelangelo, Rembrandt, Claude Monet are some of the legendary artists of all time. For instance, Leonardo is known for his dramatic and expressive masterpieces like the Mona Lisa, while Rembrandt is known for paintings that evoke the innermost feelings of his subjects through facial expressions and dramatic use of light and shadow. Each of these artists and the other famous ones, although they worked with similar mediums, developed their own signature style, refined over the years through experience, training, gathering information and honing their skills. And what distinguishes one artist from another is their unique and hard-to-replicate expression. Artists and learners have spent years attempting to replicate renowned artists known for their unique styles. The Mona Lisa, for instance, has been replicated countless times by Leonardo's students and contemporaries, yet it remains Leonardo's signature piece.

Those who go down in history are the ones who have found that unique expression. That is what every artist, new or old, aspires to achieve.

Why am I suddenly talking about art in a book on data? Well, art is one of the most complex manifestations of translating tacit

knowledge—insights, skills and abilities that an individual gains through experiences that are often difficult to put into words or some explicit form—a form that can be readily articulated, codified, stored and accessed. And the signature style that an artist develops is the consequence of years spent accumulating knowledge and expertise in their field coupled with their own innate talent which eventually becomes proprietary. Also, this proprietary aspect keeps evolving with the passage of time as well.

In the Big Data world, organizations have access to multiple sources, multiple types of data sets like I discussed in Chapter 8, Multi-Source Data and Chapter 9, Real-Time Data. Now when data is pouring in from all directions and many organizations are leveraging it to enhance their business models, the question arises that how can organizations differentiate themselves in a world where most of the data is openly and abundantly available to all?

So, let me introduce you to the third hero of our data story—proprietary data. Something unique that can help organizations stand out in a crowd. And in this chapter, instead of just talking about data, I am going to talk about a broader concept which is proprietary knowledge.

Proprietary knowledge: An organization's Mona Lisa

Proprietary knowledge is the accumulation of differentiated knowledge assets developed by codifying tacit knowledge generated across the organization's data value chain as well as methods, processes, and techniques unique to the organization, which are built over time and can be the source of durable competitive advantage.

In this data-abundant world, most organizations have access to data similar to everyone else. And as more and more organizations are implementing advanced technologies in their operations, most organizations are feeding their models with abundant internal data and openly available external data sets. It is therefore becoming harder to differentiate when most of the data is available to all. In such a scenario, proprietary knowledge can become that unique

differentiator that helps organizations set themselves apart from their competitors.

Proprietary knowledge can be created at any point along the data management value chain. At each of these stages, organizations may develop proprietary knowledge that gives them a competitive advantage. For example, at the data-collection stage, an organization may identify some proprietary sources of data that can help them derive deeper and differentiated insights. At the insight generation stage, an organization may use various methods, techniques or models that enable it to generate unique insights for better decision-making or drive more effective actions. Or at the action stage, where an organization may be able to devise better strategies, new approaches to problem-solving, or develop proprietary software or tools that enable organizations to engage the end users in new and innovative ways. As a result, organizations can significantly improve their operations or deliver superior customer experiences.

Regardless of where proprietary knowledge is created along the data value chain, it is important for organizations to find ways to capture and codify this knowledge and make it available for use across the organization. Because it is this unique knowledge base created by the organization over time that can be leveraged continuously to create and maintain differentiation in the market. What makes it even more unique is that it is not easily replicable by the competitors either.

Another aspect of proprietary knowledge is that this knowledge is often available in the form of tacit knowledge at an individual expert level or at a team level. It is the unique experiences of individuals and teams in the organization that helps them recognize patterns over time and connect them in ways to deliver exceptional value. The challenge lies in capturing this tacit knowledge to make it available in explicit form for consumption on a continuous basis in the organization.

According to a survey, 97 per cent of the companies surveyed acknowledged that proprietary data was 'very valuable' or 'quite valuable' for differentiating one's company from competitors. The

same survey also revealed that only 10 per cent of respondent considered their company's proprietary data more useful than the competition's.¹

This highlights both the importance of proprietary knowledge as well as the challenge of creating it.

Proprietary Data is the key source of differentiation for organizations



But when done right, it can unleash tremendous value for the organization, enabling it to achieve a sustainable competitive advantage over their competitors. Just like salt, although added in a very small quantity, it is essential to bring out the flavour of the soup. Similarly, proprietary knowledge, which is unique to the organization, though relatively less in quantity, acts as that essential pinch of salt that brings out the flavour of differentiation essential for sustainable competitive advantage.

My close encounter with proprietary knowledge

This topic is very close to my heart because I have been fortunate enough to be centrally involved in conceptualizing and driving the journey of building proprietary knowledge at McKinsey, during the

course of setting up the McKC, McKinsey Knowledge Center. What evades many people is that one of the key foundations on which McKinsey, one of the premier management consulting firms in the world, was built, was proprietary knowledge. It truly differentiates McKinsey, leading it to its sustained pre-eminence. So let me take you through the story of how it happened.

Of course, it wasn't something we achieved on day one; it was a journey, a process of growing and evolving to put together our own Mona Lisa. From the very beginning, McKinsey had a deeply ingrained process and culture of building knowledge assets. Whenever a client engagement was completed, consultants were encouraged to document it as practice documents (PDs), to translate them into shareable insights. These documents were stored in one place, like a repository, where it was accessible by everyone in the firm. And owners of PDs that were used more frequently than the others were then recognized and rewarded, reinforcing it as part of McKinsey culture.

Then when we started developing McKC, we saw this as an opportunity to take the knowledge centre beyond a run-of-the-mill back-end support function. We foresaw the value of converting the PDs, which were typically in the form of PowerPoint or Word documents, difficult to scale and very difficult to keep updated. By creating databases which captured the key data and insights from it, we transformed these into standardized formats. That was the turning point. The database grew continuously due to an efficient format that could be easily and regularly updated. Unlike traditional consultant-developed PDs, specialized teams maintained our databases, creating a readily accessible proprietary knowledge base. This scalable model allowed us to capture and maintain more proprietary information, enriching McKinsey's knowledge resources.

Now let me tell you how it became the source of competitive advantage for McKinsey. I was a diagnostic expert at McKinsey from the beginning. My role, as I also discussed in the introductory chapter, was to run benchmarking diagnostics, which were often a precursor for our client engagements. Since we did not have well-established benchmarking standards at that time, we used to come

up with our own set of KPIs and key benchmarking analyses based on certain educated assumptions and expert advice. The issue was that for every new client engagement, this process ended up happening in an ad hoc manner. But when I moved from consulting to the knowledge centre, I realized the potential of bringing structure and scalability to this process. We started capturing key data being generated by every diagnostic being performed across the firm. With time, we ended up creating a solid database that was then leveraged to create comprehensive benchmarks for diagnostics that were customized for various industries or functions, like tech services, manufacturing, supply-chain procurement, etc. This process involved combining both experiential knowledge, which included our proprietary benchmarks and frameworks, with external data to create our own proprietary diagnostics.

So, the process of codifying the data and analysis around each customer engagement helped create our own substantial proprietary knowledge base. Additionally, this enabled us to create a virtuous cycle; whereas more and more data started getting accumulated, we were able to generate deeper insights—for example, we could do more correlations, identify causal factors more accurately, etc., and for the first time, we could arrive at statistically relevant insights. Now we started to offer benchmarking as a standalone client service, which increased the scale from a few clients to a substantial number. And as we continued to enrich our database, we were able to further refine and improve our diagnostic model.

Over a period of time as external data started to explode, industry benchmarks that were proprietary in the past were increasingly being made available externally. So McKinsey also started combining external data sets as well with the internally generated benchmarks to refine the process and models further. Now just relying on data alone wasn't enough to sustain our competitive advantage, because competitors had also built their own proprietary databases over the years. So, the very definition of proprietary knowledge needed to evolve. Now McKinsey was able to maintain their differentiation through their knowledge and understanding of critical KPIs across

specific industries and functional areas because of the work done by them in these areas over years.

As highlighted in Chapter 7 on the importance of breaking down the business problem to identify the most critical KPIs, McKinsey has built expertise in putting together an effective framework of the few very critical input and output KPIs that would help provide the sharpest insights or achieve the maximum impact. And that has now become its unique differentiator. So now McKinsey is in a position to provide benchmarks customized to most industries and functional topics. Most consulting engagements start with a diagnostic, supported by these custom-built benchmarks, enabling McKinsey to build a durable competitive advantage over its competitors. So much so that these diagnostics have become the signature style for McKinsey, its 'Mona Lisa'. What started with tacit knowledge with the consultants, turned into benchmarks over a period of time, then into diagnostics that included external and internal data, then grew bigger than the data, into building the framework around critical input and output KPIs, which over a period of time could be delivered directly to clients through our platforms. Therefore, the very definition of what is proprietary knowledge kept evolving to more than just the data.

The tacit nature of proprietary knowledge

The core of building proprietary knowledge lies in the practical insights of those who've worked within the organization. They possess valuable understanding of the business, operations, products and processes. This evolving knowledge, derived from data, methods and techniques, can enhance efficiency and scalability, offering a lasting competitive edge. For example, a manufacturing company with unique processes for efficiency and waste-reduction not only achieves immediate cost savings but also pioneers sustainable operations, curbing its carbon footprint in the long run.

So, proprietary knowledge comes from the experiences gained by organizations or individuals in their work. Think of a skilled

salesperson who can tell apart customers likely to buy from those who won't, just by observing their behaviour. For any organization this valuable insight can help predict sales and improve customer engagement, similar to how online stores analyse user behaviour. But for physical retailers, this wisdom often remains unused in the salesperson's experience and is not codified. Such knowledge might get passed through apprenticeship and even interactions but is typically not available widely across the organization. During my tenure at McKinsey, we believed that the majority of the organization's knowledge exists in tacit form. And despite all the technology advancements, even today it holds true, with more than 80 per cent of the organization's knowledge being tacit in nature.²

This raises the question: How can we capture such valuable tacit knowledge?

Key to unleashing the power of proprietary knowledge —‘codify’

Since proprietary knowledge is mostly generated through the work done, embedded in the practices and routines of employees, it is often subjective, informal. It is difficult to share or express, or at times even identify such knowledge, and therefore requires a focused effort to capture it in explicit form. Another issue with such tacit knowledge is the risk of losing it when an experienced employee leaves the organization, especially if the employee was the only one with that specific expertise or skill set. And since most organizations do not have a structured process to capture the tacit knowledge, it often goes untapped.

So the key is to have a structured approach to ‘codify’ the tacit knowledge so that it can be made available for use broadly and on-demand. Codifying tacit knowledge means making it explicit and storing it in a structured and organized format, which can be easily accessed and shared across the organization.

But the process of codifying this knowledge is not an easy one. The biggest challenge in codifying tacit knowledge is that individual

experiences and insights are often highly specific, making it hard to translate this information into a more broadly applicable format. In addition, unless this process is embedded in the organization's culture, it is an added burden on the individual to capture and document their knowledge. It falls into the category of those not urgent but important tasks that require an extra effort from the employees. Therefore, it is critical for organizations to formulate a well-thought-out plan to capture and document this knowledge on an ongoing basis through knowledge-management processes and systems.

Codifying tacit knowledge requires a fundamental shift

In a recent survey, while 75 per cent said that creating and preserving knowledge across evolving workforces was important for their success, only 9 per cent agreed that they are ready to address this.³

Though crucial, implementing an organization-wide strategy for making proprietary knowledge a centrepiece isn't easy. It needs a fundamental shift in how core business functions gather, develop and share knowledge. Knowledge management must integrate into the business strategy through a structured approach. This involves establishing an effective knowledge cycle, aligning assets with strategic goals. A senior leader at McKinsey once said, 'Knowledge is McKinsey's lifeblood ... We've invested more in knowledge recently. If that means 10–15 per cent less client work now, we'll invest in the future.' As explained earlier, McKinsey's competitive edge comes from a well-structured, enriched knowledge cycle developed over years.

Knowledge marketplace: An approach to create a virtuous cycle of knowledge

Building a sustainable knowledge management practice involves building a virtuous cycle of knowledge—at the heart of it is creating

a knowledge marketplace operating on the principle of supply and demand matching.

The knowledge marketplace approach is based on developing organization-wide capability where the demand of knowledge seekers is catered to by the supply from knowledge contributors. The seekers and contributors come together to build and sustain a self-creating self-perpetuating virtuous cycle of knowledge. The marketplace is a way to bring together both explicit and tacit knowledge in a codified form available to be accessed by anyone in the firm.

The knowledge of the organization comprises various assets like expert knowledge, codified documents, data models, etc., that are codified to be made available in the knowledge market for the knowledge seekers like employees, team members to be leveraged as and when required. The knowledge brokers are the guardians responsible to curate and maintain the knowledge systems like repositories, directories, etc., to keep them up-to-date and make them easy to track and access, enabling these assets to be used in a repeatable manner. For instance, take Kaggle's knowledge marketplace for machine learning. It connects knowledge providers with knowledge seekers through a vast array of organized and easily accessible information on topics related to data science, ML and AI. Users can search by keyword, topic or skill level and rate and review content to help other knowledge seekers. It has evolved into a thriving community where data scientists and machine-learning practitioners can freely exchange and sell their data science and ML expertise, like models, data sets and tutorials.

There are six levers essential for unlocking the power of the knowledge cycle. I call them the knowledge-management pyramid.

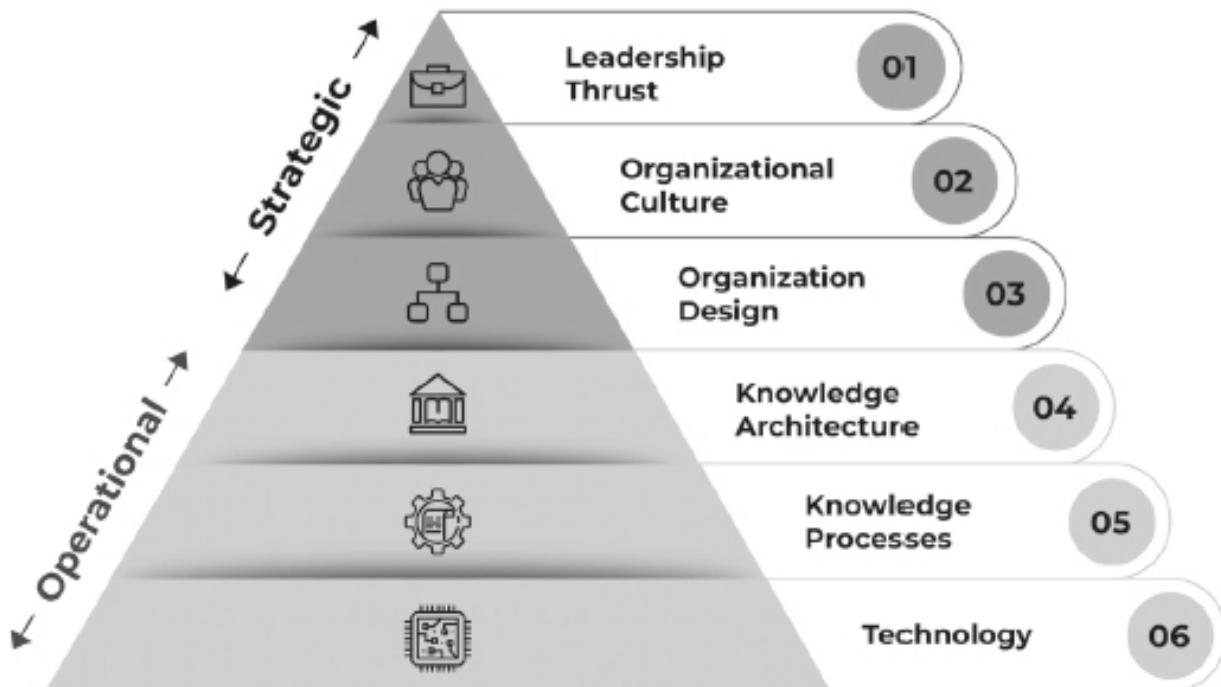
Strategic levers

- 1. Leadership thrust:** Leadership commitment and repeated reinforcement of the criticality of knowledge to the organization is crucial for the success of any

knowledge-management initiative. Explicitly embedding knowledge aspirations in the company's strategy and values is important to build a knowledge culture.

2. **Organizational culture:** Creating a collaborative and inclusive culture that fosters the exchange of knowledge from both internal and external sources, while incentivizing participation and recognizing its impact, is of equal significance.
3. **Organizational design:** Establishing a sustainable practice in knowledge management also involves integrating knowledge management deliverables into individual roles and responsibilities, linking them to KPIs, and creating dedicated roles within the organization to oversee the knowledge management cycle (as we saw in the McKC example).

Knowledge management pyramid



Operational levers

1. **Knowledge architecture:** Building a knowledge map that acts as a blueprint for a well-designed knowledge architecture is essential to enable superior knowledge discovery, that is embedded with superior search capabilities beyond classic keyword searches to a more cognitive or intuitive search that can bring relevant content together from multiple knowledge sources.
2. **Knowledge processes:** Putting in place the right processes and methods for content creation and curation based on the existing bodies of work and enforcing a systematic process for knowledge capture, standardization and availability. Supporting processes like taxonomy management, audit mechanism, etc., is critical to ensure consistency, quality and relevance at all times. It is also critical to ensure that the knowledge asset is up to date and of high value to the users.
3. **Technology:** Emerging digital technologies like AI/ML, natural language processing (NLP), offer new and distinct capabilities that can make the entire process more efficient and scalable. For example, Incedo has implemented an AI-based knowledge management platform which streamlines the entire process. First, it gathers knowledge from diverse sources and deploys theme/domain-based categorization on the knowledge assets. It then extracts and packages the insights from the knowledge accumulated. And finally, it manages the delivery of the knowledge to users by optimizing the experience to specific requirements and preferences.

The growing importance of algorithms

The process of creating proprietary knowledge is a journey. And while organizations continue to build their proprietary assets, the very nature of what is proprietary is also constantly evolving. While

data, insights, methods and processes continue to contribute significantly to what is proprietary, now algorithms are becoming disproportionately important in the AI age. The use of data through unique algorithms are increasingly providing organizations with critical capabilities to redesign their operations, develop innovative products and services, and gain a sustainable competitive edge in the market. Netflix is a great example of innovative use of data and algorithms to create differentiated offerings. They collect viewer data which includes device details, date, time, location, duration of a particular content watched, data about searches on its platform, portions of content re-watched, and its frequency, and even data about content being abandoned. Additionally, they also collect metadata from third parties such as Nielsen and data from social media sources like Facebook, X (formerly Twitter), etc. Leveraging these data points on the interaction of millions of subscribers with their content, their unique algorithms generate insights that are used to create original content to meet the unique preferences of their customers.

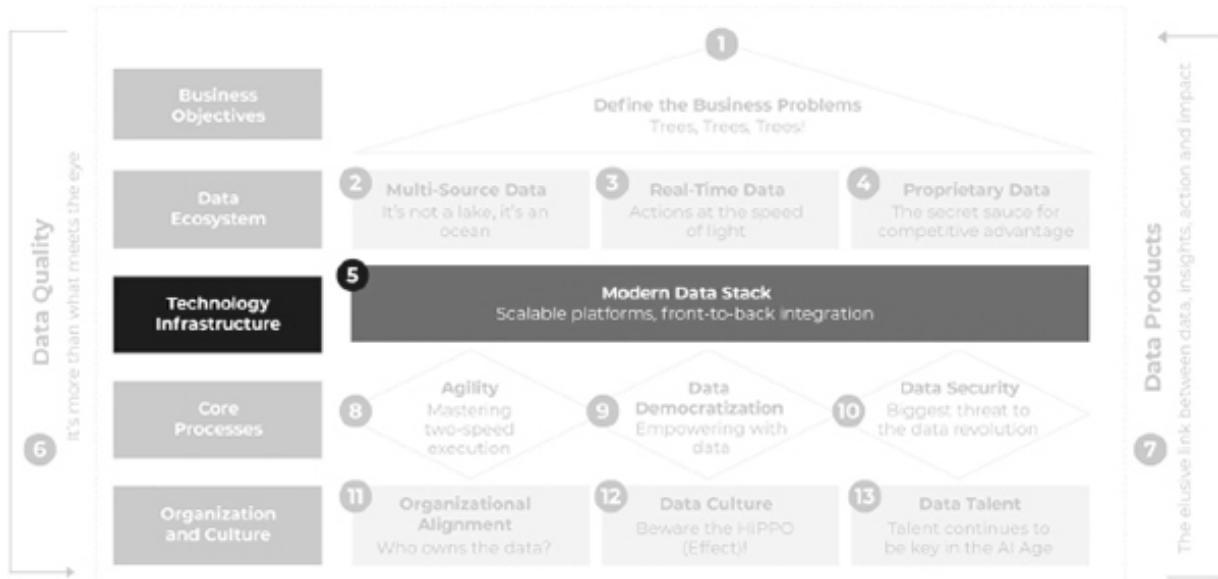
One of the most noteworthy utilizations of these proprietary insights is the success story of a highly popular series created by Netflix—*House of Cards*. Through their proprietary algorithms they identified some critical success factors such as correlation between the star, the director and popularity of the content featuring them. Based on this they concluded that producing a series for the American audience, starring much-liked actor Kevin Spacey from a successful British show, and director of the already popular series *The Social Network*, is bound to be a big hit. According to Netflix, *House of Cards* was such a success that it was the most streamed piece of content in the United States and forty additional countries at the height of its success. And through the years, it continues to be one of the top-rated series on Netflix, putting it in the league of blockbusters like *Avatar* and *The Sopranos*.

Key takeaways

- In the Big Data world, where data is abundantly available to organizations, it is difficult to maintain sustainable differentiation based on data alone. Proprietary knowledge, the accumulation of differentiated digital assets and codified tacit knowledge, can offer a lasting edge.
- Proprietary knowledge is still largely tacit, therefore codifying knowledge becomes essential. This is a fundamental shift, and knowledge management needs to become an integral part of an organization's business strategy and culture.
- A structured approach is needed to build a virtuous cycle of knowledge. The key to it is enabling a knowledge marketplace where knowledge seekers' demands are met by knowledge contributors' supply.

LAYER 3

TECHNOLOGY INFRASTRUCTURE



'The foundation to bring it all together'

11

Modern Data Stack

Scalable Platforms, Front-to-Back Integration

'May you have a strong foundation when the winds of change shift ... and may you be forever young.'

—Bob Dylan,
American singer-songwriter

The entire technology for creating value from data has gone through a structural upheaval. The traditional approach of dealing with data, designed for limited, mostly structured data in batches, has failed to provide an effective solution to the demands of the Big Data world. According to a global survey, one of the top three challenges companies face in meeting their analytics objectives, is designing data architecture and technology infrastructure that can effectively enables data and analytics at scale.¹ Therefore, it is critical to redesign the data architecture, to make it more **scalable, agile and adaptable** to support the continuously evolving nature of data and the changing needs of business users—to seamlessly manage the winds of change and stay relevant. This is why the data architecture or the data stack—how data is collected, transformed and used, must be 'modernized'.

What is a data stack?

A data stack is like the foundation, the structure critical for organizations to be able to leverage and manage data effectively and in a timely manner. But what exactly is it? A lot of attempts have been made to define a data stack or what makes it modern. And

arguably it is one of the most difficult concepts of this book to define or explain because it is not just one thing. It's a combination of technology capabilities needed to translate raw data into the finished product—decisions and actions. So let me attempt to simplify this concept as much as possible without getting too much into technical details.

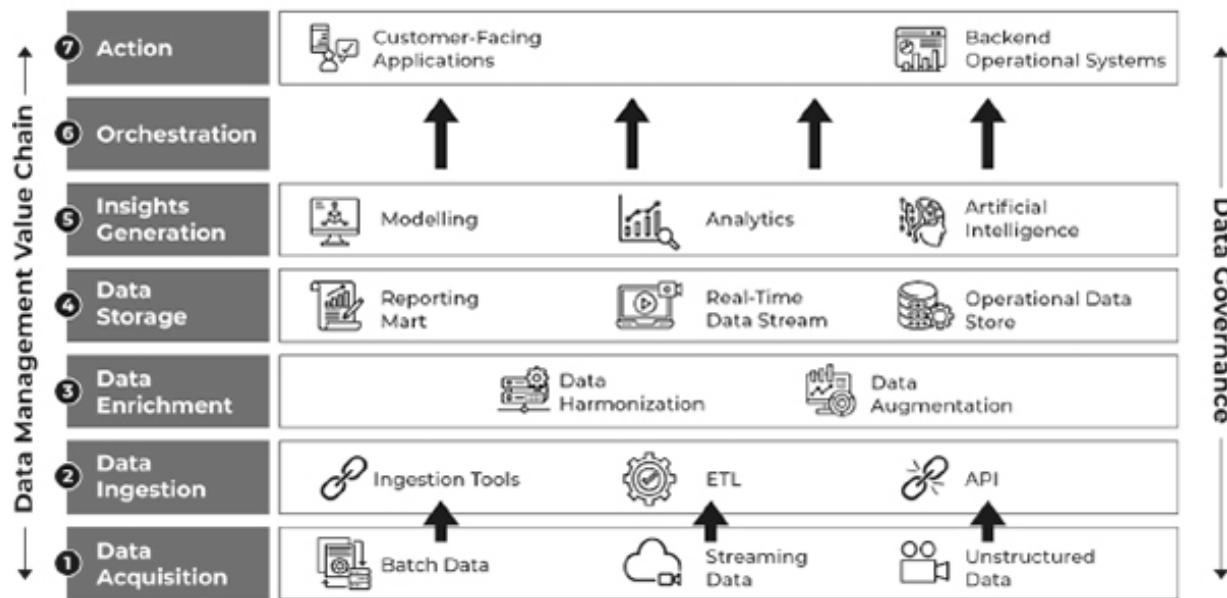
A data stack is the combination of multiple tools, technologies and processes that enable companies to leverage data for decision-making. It includes tools/technologies that are used for collecting, processing, storing, analysing and driving actions from data.

So, as the definition suggests, it enables every layer of the data management value chain that we talked about in Chapter 3, Value Reimagined. To make it easier to understand, let me explain the multiple layers that make up the data stack using a kitchen analogy. Think of a modern data stack as a kitchen where you prepare a meal. You have all the necessary ingredients, appliances and utensils at your disposal to cook and consume the meal. Each layer is like a step towards getting the meal ready and finally consuming it:

Step 1: Data acquisition—gathering the ingredients

Data acquisition is the process of collecting or gathering data from different sources. It involves the capture of raw data from sensors, devices, databases, APIs, or other external systems from streaming or batch processes. It is synonymous to gathering or shopping for ingredients of different kinds like fruits, vegetables, spices, etc., from different sources or vendors to bring into the kitchen.

What is a Modern Data Stack?



Step 2: Data ingestion—bring the ingredients in

Then we bring in and organize the ingredients in the kitchen. Data ingestion is the process of taking acquired data and loading it into a target system for storage, processing or analysis. It involves preparing, transforming and loading the acquired data into a suitable data storage or processing platform, such as a database, data warehouse or data lake.

Step 3: Data enrichment—seasoning for flavour

Like adding seasoning to enhance the flavour of a dish, data enrichment is the process of enhancing or augmenting existing data with additional information or attributes to make it more valuable and useful for analysis, decision-making or other purposes. It involves enhancing existing data sets, thereby improving its quality, depth and comprehensiveness.

Step 4: Data storage—choosing the right utensil

In a kitchen, a cooking utensil is chosen based on the ingredient you want to cook. Similarly, data-storage platforms are selected primarily based on the nature of data involved in various use cases. This can include traditional databases like relational databases where collection of data items is done with predefined relationships between them, or object-oriented databases in which information is represented in the form of objects, as well as newer options for handling Big Data like NoSQL databases which are non-tabular databases, which store data differently than traditional databases.

Step 5: Insights generation—the cooking process

Just as you use various appliances like the stove, microwave or other appliances where the cooking utensils are loaded to transform ingredients into desired dishes, insight generation involves using data for analytics, data modelling, business intelligence and operational analytics. This layer helps derive relevant insights to improve decision-making or support advanced AI/ML use cases.

Step 6: Orchestration—the mechanics of serving

Orchestration is like serving the cooked food in appropriate utensils like plates or bowls, etc., to make it easy for people to consume. Orchestration refers to the process of translating insights into actions. It involves the coordination and automation of various tasks, workflows and actions to drive real-world impact based on the insights derived from data.

Step 7: Consumption—bon appétit

Finally, the various dishes cooked and served are consumed. The consumption layer is where the insights are leveraged both internally by decision-makers or externally by end customers to take data-driven decisions or drive actions. This can be through internal operational and business intelligence systems by business leaders or

customer-facing portals, applications or any other system of engagement. This is the final layer that helps translate insights into actions like decisions around day-to-day business operations, customizing user experience, making process improvements, incorporating product enhancements, etc.

Why is the traditional data stack (TDS) not enough anymore?

Traditional data stacks (TDS) are typically built as on-premises solutions, where the data storage infrastructure is either owned and managed by the organization or outsourced to third-party service providers. They are primarily designed as monolithic architectures (where all components and functionalities are tightly integrated into a single, unified system) and are difficult to integrate into cloud-based environments. As a result, these are limited in their flexibility and scalability which is essential for businesses operating in today's dynamic environment. But why has the situation drastically changed in recent years? It is because, across each stage of the DIAI framework, both the nature of input (that is, the data) and the expected outcome have evolved significantly. This is putting tremendous pressure on the traditional data stack. I have spoken about it in parts in various chapters in Section I and II so far. Let me bring it together here to emphasize on how?

1. **The 3V nature of data is putting pressure on the TDS**

TDS built as on-prem infrastructures struggle with efficiently managing large data volumes, leading to overwhelmed storage, processing and analysis capability that cause performance bottlenecks and slower responses. These stacks often rely on structured storage systems, like relational databases, which have limitations in handling extensive data, resulting in storage constraints, hindered data access and slower execution of tasks like querying and analytics. Additionally, the fast

pace of real-time data is a challenge for TDS as they lack the capacity to process such data promptly, leading to latency issues and outdated insights. The high data velocity puts a strain on TDS due to scalability constraints, hindering their ability to handle sudden data surges. TDS also struggles with the varied formats of data, which require specialized tools for integration and processing data, further highlighting their limitations in integrating and harmonizing diverse data sources.

2. Insight generation has to be high velocity and in real time

Organizations operating in a complex ecosystem amidst dynamic business environments require a more proactive and real-time approach to problem-solving. To do so effectively, they are moving towards prescriptive analytics from merely descriptive ones, to enable a more proactive and forward-looking decision-making. Furthermore, to make high-velocity decisions and take actions in real time, organizations are shifting from batch processing to real time, where analytics is becoming a continuous process. Additionally, organizations are increasingly leveraging AI to automate and enhance the depth and breadth of insights generated and speed up the whole process. TDS are designed for insight generation on structured data and leverages batch processing which fails to support high-volume, high-velocity insight generation in real-time.

3. Action or decision-making must be instantaneous and automated

In today's high-velocity decision-making landscape, time is of the essence. Actions need to be instantaneous and automated, which requires data to instantly transform into insights and drive real-time actions. Additionally, with AI playing a significant role in automating insights to

action, the TDS typically lack the automation and orchestration capabilities to enable such automated actions. TDS also lack the capabilities that can enable swift decentralized data-driven decision-making that can empower employees at all levels to make informed choices using seamless self-serve data capabilities.

4. Transformational impact from data

Data's transformative potential, as explored in the initial chapters, offers organizations the means to create innovative business models, enabling exceptional customer experiences, operational enhancements and innovative differentiation, as exemplified by disruptors like Uber, Airbnb and Netflix. The extensive data ecosystem also empowers organizations to monetize data as an asset, whether by selling it to third parties or offering tailored products and services, such as targeted advertising on social media platforms or packaged insights to businesses by research or consulting firms. Furthermore, in the digital age of dynamic, constantly evolving business environments, organizations must leverage data to create a virtuous cycle of learning and improvement. To do so effectively, organizations must leverage data from diverse sources to generate deeper and more meaningful insights, innovate at scale and provide improved customer experiences on a continuous basis. TDS, with its limitations around real-time processing, scalability, flexibility, advanced analytics, integration capabilities and slow iteration cycles are ill-suited for fully harnessing the transformative potential of data.

A recent survey suggests that more than half of the respondents still find it challenging to leverage data to provide new products and services to gain a competitive edge (57 per cent), adapt products and services to meet market needs (55 per cent), and deliver data-driven

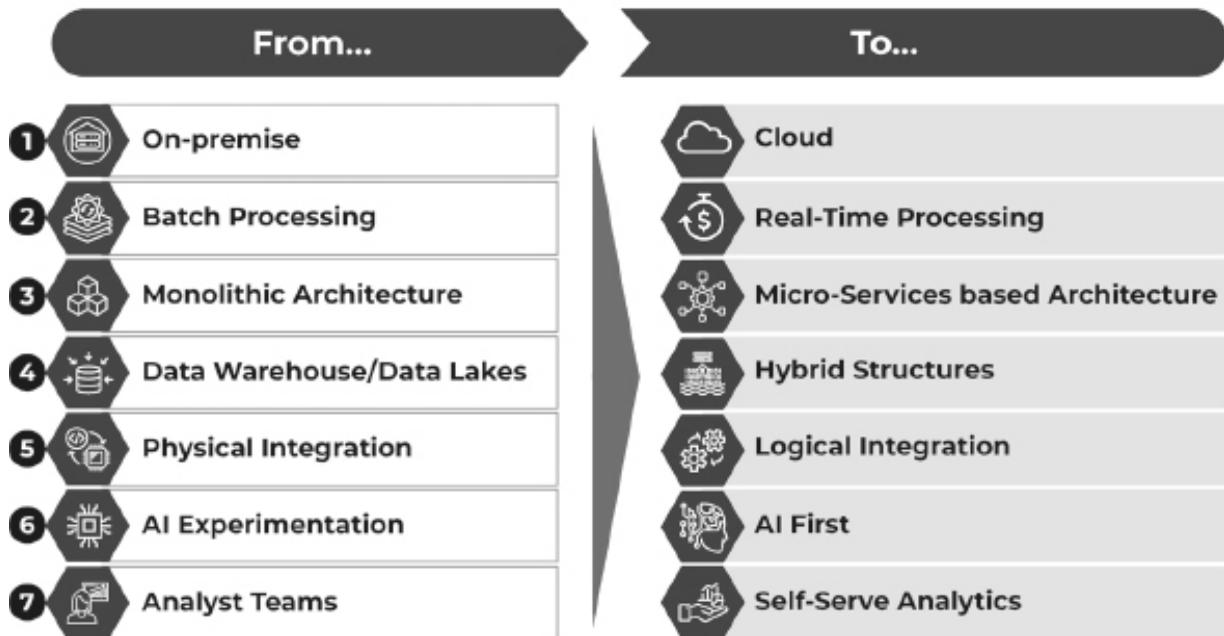
decisions for businesses (52 per cent).² So, leveraging data effectively is increasingly becoming a critical priority, for which organizations require an agile, scalable data architecture, which TDS lack.

What's so 'modern' about the modern data stack (MDS)?

A modern data stack is a cloud-hosted, comprehensive set of tools and technologies, where the different layers of the data management value chain are loosely coupled yet tightly connected, for rapid scaling, adaptability and seamless interoperability, to drive the data to insights to actions to impact cycle, at speed.

Building a modern data stack would not have been possible without some key technological shifts that have happened over the past few years. Newer tools and technologies that have emerged in the recent past can help tackle the complex nature of Big Data, and enable organizations to build a scalable, elastic and adaptable modern data stack. Let me highlight some of these key technology shifts:

Key technology shifts enabling modern data stack



I have talked in detail about a few of these shifts in earlier chapters. In this section, I will elaborate more on those shifts that haven't been covered elsewhere in the book.

1. From on-premise to cloud

The most disruptive shift that has enabled a fundamental transformation in the data infrastructure is cloud. The traditional on-premises IT infrastructure posed a challenge in terms of lack of elasticity and scalability, and needless to say, was costly to install and maintain. Enter 'Cloud'. Although AWS had launched its modern cloud infrastructure service with the Amazon Elastic Compute cloud in 2006, the pivotal shift happened with Amazon Redshift, its data-warehousing cloud solution, launched in 2012, paving the way for modern data stack.³ It gave organizations the choice to build an asset-light data infrastructure that is owned and managed by third parties, enabling them to focus on their core businesses.

Over time, the cloud has expanded beyond basic storage and computing. It began with Infrastructure as a Service (IaaS), giving

users control over infrastructure components, like virtual machines, storage and networks. It further evolved to Platform as a Service (PaaS), offering a ready-to-use cloud-hosted platform to build, deploy, run and manage applications without worrying about the underlying infrastructure. Finally, Software as a Service (SaaS) expanded it further by delivering apps over the internet, eliminating local installations and maintenance requirements. The underlying infrastructure, platform and application code are managed by the SaaS provider. Organizations can now pick cloud models to best suit their needs.

Owing to this convenience of renting storage and computing capabilities on a pay-per-use basis, companies have increasingly been outsourcing their infrastructure to external cloud providers or the 'hyperscalers'—*AWS*, *Microsoft*, *Google*, *Alibaba* and *Oracle*, among others.

Role of hyperscalers

Hyper means excessive and scale means the ability to increase or decrease in size, number or extent. So, the word 'hyperscale' refers to the ability of a system or architecture to scale as much as required and whenever required. The hyperscalers—large cloud-service providers who offer massive storage and computing resources, typically in the form of an elastic cloud platform, are the ones that make it possible. This trend was started by Google in the 2000s, which opted for a free and open-source operating system called Linux, as it realized its search-engine business could not be viable in the long term if it depends on its own traditional high-end servers. And then AWS followed suit and so on. Today hyperscalers offer such a comprehensive suite of services and have brought together an entire ecosystem of service providers, making it easier for organizations to build a modern data stack.

Here is how hyperscalers have played a significant role in enabling and driving the adoption of modern data stacks:

- **Infrastructure for scalability:** Cloud-based infrastructure offers limitless scalability and elasticity on demand. Hyperscalers offer highly scalable infrastructure resources, including computing, storage and networking capabilities. Businesses can easily adjust capacity as per their need: they can scale vertically (adding more resources such as RAM or processing power to a server), horizontally (adding servers to distribute workload) and diagonally (a combination of both). This flexibility allows handling of large volumes of data and accommodates spikes without major infrastructure investments. Deploying new applications and services is easier and faster, without hardware or software purchases. Features like automatic backups, replication and disaster-recovery ensure data durability and high availability, reducing the complexity and upfront investments needed for data management.
- **Secure and reliable services:** Cloud-based data infrastructures are more secure and reliable than the on-prem ones. Although counter-intuitive, since organizations share virtual space, these service providers adhere to the highest security standards, including international certifications and third-party audits and validations, making them more secure. The cloud infrastructure is made more reliable by a team of experts responsible for data integrity, ensuring uninterrupted, secure services, even in disaster scenarios. Hyperscalers prioritize data security, with robust measures like encryption, access controls and compliance certifications. They help organizations meet regulatory requirements in various industries, for secure data management and processing.
- **Pay-per-use:** In a cloud-based data stack—unlike on-prem technology infrastructure that requires huge investment to establish and maintain—organizations use storage and software applications hosted on remote

servers owned and maintained by the cloud service providers. This eliminates the need to invest in data centres, servers, physical networking or maintenance activities; companies are billed only for actual usage of resources like storage, computing, etc. So, they avoid large, fixed investments while getting access to a wide variety of managed services from their cloud service providers like IaaS, PaaS or SaaS.

- **Ecosystem and integration:** Hyperscalers provide a solid foundation for a vast ecosystem of third-party service providers to offer compatible services, tools and integrations that complement and extend the capabilities of the modern data stack. This ecosystem includes data-visualization tools, data-governance solutions, data-cataloguing platforms and more. The plug-and-play solutions can easily be integrated to customize the data stack as per the organization's business needs. They continue to offer newer and better capabilities for organizations.
- **Enhanced democratization:** Cloud-based data infrastructure also enables easy accessibility and better collaboration. The hyperscalers—by providing access to a wide range of self-service analytics tools, scalability for large-scale data processing, collaboration and sharing features, and robust data governance and security—empower users from various roles and departments to easily access and analyse data, make informed decisions, and collaborate on data-driven projects. Cloud-based analytical tools are more powerful and accessible anywhere, anytime for business users across all levels in the organization.

Overall, building a modern data stack before the advent of hyperscalers was a complex and resource-intensive undertaking that required significant expertise in infrastructure management, data storage, processing, analytics and integration. Hyperscalers have

alleviated many of the challenges and simplified the implementation and management of modern data stacks.

2. Batch to real-time processing

With the growing proportion of data being generated and consumed in real-time, it is critical for a modern data stack to enable the process of generating insights and drive decision-making and actions in real time. With data-management cycles collapsing significantly, organizations are increasingly being crippled by the long turnaround time that a traditional data stack is unable to cut down on.

Every layer in a traditional data stack was built for the batch world. The data ingestion from traditional sources, transforming and then storing it in on-premises data warehouse or even in the cloud and BI tools were all designed to work with structured data to build reports after the event has taken place.

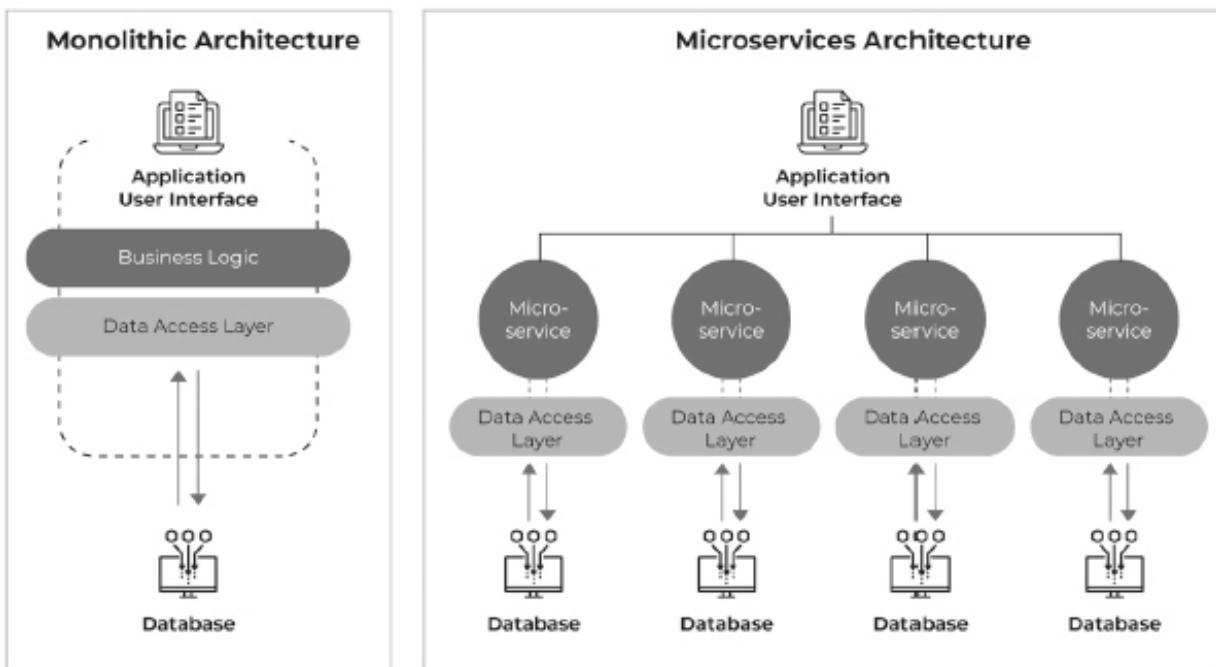
Today, the real-time data infrastructure is enabled by a new category of applications like real-time analytics and automated solutions that are being incorporated at every stage of the data stack. Apache Kafka for streaming data, processing tools like Apache Flink^{*} to filter and transform real-time data enroute, ClickHouse[†] or Tinybird[‡] to analyse it in real time or Tecton[§] based on a machine-learning model to continuously adapt and learn from data and generate predictions on-the-go, are a few examples of real-time applications or tools that have enabled the modern data stack to deal with real-time data effectively. Netflix over the years has built a real-time data infrastructure that probably handles trillions of events every day. Netflix has managed to expand the utilization of real-time data from zero to thousands of use cases in the past few years. It has been able to do so by building capabilities along various layers of the data stack.⁴

Details on real-time data and analytics have been discussed in Chapter 9, Real-Time Data.

3. Monolithic to microservices architecture

Monolith refers to something that is built or composed as one. A monolithic approach to building a data stack involves designing and implementing the entire stack as a single, cohesive unit. In this approach, all components of the data stack, such as databases, data-processing and analytics, etc., are tightly integrated and bundled together within a single application or system.

Monolithic Architecture vs Microservices Architecture



While monolithic architectures are easier to build and deploy, they have certain drawbacks that make them unsuitable for bigger and more complex deployments. These architectures are suitable for small-scale implementation because they are built as a single unit implemented for fewer use cases. But as the volume and complexity of data increases, organizations need to use multiple, varied data sets to enable larger and more complex use cases. In such a scenario, monolith becomes a constraint due to its lack of scalability, flexibility and modularity. Tightly coupled architecture also makes it very difficult to iterate, because updates or changes to any component in the data stack typically require redeploying the entire

monolith. In addition to that, resources such as computing power, memory and storage are shared among all components within the monolith. Scaling individual components can be challenging and often requires scaling the entire stack.

Organizations are rapidly recognizing the value of a flexible yet interconnected structure of a microservices-based architecture—a collection of small, independent services that can be developed, deployed and scaled independently, an approach that successful tech giants like Amazon, Google and Microsoft have built their offerings on.

In the context of data stack, microservices-based architecture is often chosen for larger and more complex data stacks that require scalability, flexibility and independent development of different components. It involves breaking down the data stack into smaller, independent services called microservices. Each microservice handles a specific function or component of the data stack and communicates with other microservices through well-defined APIs (application programming interfaces are like contracts between two applications specifying how the two would communicate).

Let me simplify this with an analogy. Monolithic architecture is like a bird's nest. It is a single, close-knit structure, where different materials like twigs, leaves, feathers, etc., are tightly woven together to form the bird's home. The entire nest functions as a single unit and any upgrading, expansion or damage to one part would require rebuilding the entire nest or building another one separately. In contrast, a microservices architecture can be compared with a honeycomb. It is made up of multiple hexagonal compartments, where each individual cell acts as a self-contained unit designated as a home for a bee to store honey or raise young ones. As the population grows or more space is needed, the hive can be expanded by attaching more hexagons to the structure. Disruption to any one cell or a couple, can be rectified independently, while the rest of the structure continues to function as usual.

In addition to the modularity and the interconnected yet independent components discussed above, following are a few

additional characteristics that make a microservices-based data architecture preferable:

- **Enables distributed data management:** Instead of every component using a shared pool of data, in a microservices-based approach, data can be distributed across multiple microservices based on their type and nature, or each microservice can have its own dedicated database or utilize specialized storage systems based on use cases.
- **Reduced impact of errors or downtime:** Since microservices are built in a modular manner—fault or error in any one microservice is contained or isolated. By decoupling services and enforcing clear boundaries, the impact can be limited to the specific microservice experiencing the issue.
- **Granular monitoring:** Each microservice be monitored and controlled separately, allowing for detailed visibility into the performance and behaviour of individual components. On the other hand, a centralized monitoring system can provide a comprehensive view.

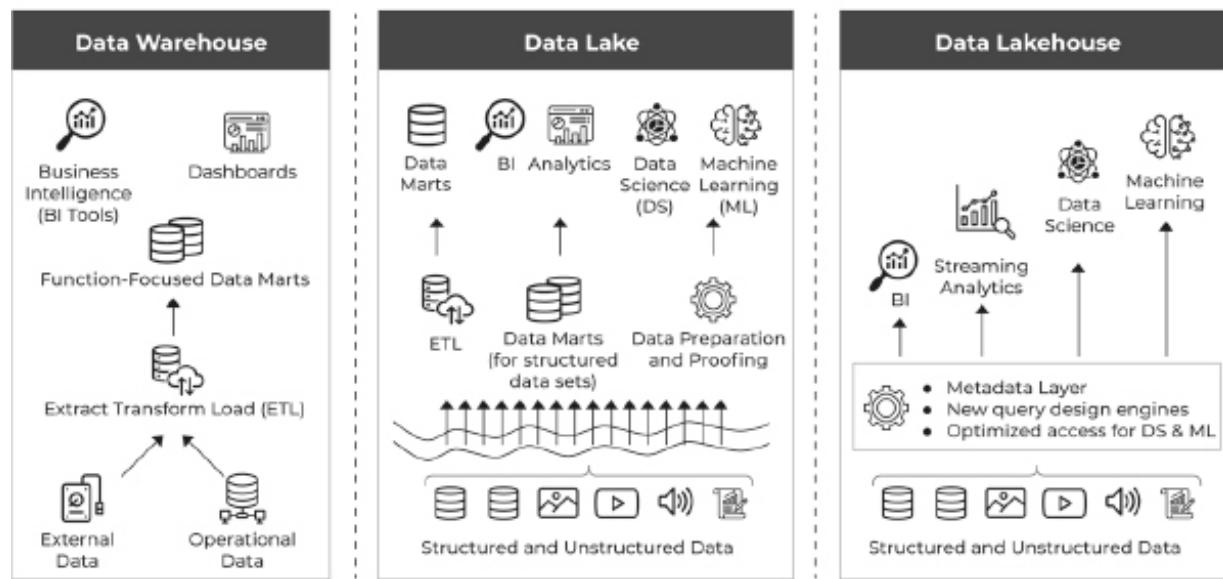
4. Data warehouse/lake to hybrid structures

Data warehousing—a concept that came into existence in the 1970s (remember Chapter 1, Data Explosion?)—was predominantly used by large corporations to store and analyse their data. But over the years, it has become more affordable and accessible, and has become a common denominator for organizations of all sizes.

Traditionally, data warehouses were hosted on-premises—often on a mainframe computer—and its functionality was focused on extracting data from other sources, cleansing and preparing the data, and loading and maintaining the data in a relational database. Now, since this concept emerged in the 1970s, while it could store unlimited data, it was designed to store it in a structured format. But as the complexity and nature of data continues to evolve with Big Data,

primarily driven by unstructured data, the traditional data warehousing approach is not scalable or agile enough to enable the dynamic nature of decision-making required by businesses today.

Data Warehouse vs Data Lake vs Data Lakehouse



The data in a traditional data warehouse is more rigid, structured and normalized, making it challenging to handle unstructured or semi-structured data such as social media feeds, log files and sensor data. Integrating data is also a challenge as data warehouses often require data to be transformed and modelled before loading, which can be time-consuming and may delay the availability of data for analysis. In addition to that, the traditional data warehousing approach also proved to be ineffective with real-time analytics as it was designed for batch processing.

So, in order to manage the unstructured, real-time data more effectively, another approach was introduced around 2011, 'data lake'. A data lake is a centralized repository that stores large volumes of raw, unprocessed and heterogeneous data. It is designed to accommodate various data types, including structured, semi-structured and unstructured data, such as text files, sensor data, log files, images, videos and more. As the name suggests, it succeeded in providing a unified storage platform where organizations could

store data, both structured and unstructured, from various sources and formats in a single location, in its raw form, often leveraging scalable and cost-effective cloud-storage solutions.

Although data lakes provide the much-needed capability to deal with unstructured, real-time data, it lacks the ease of drawing insights from structured or transactional data, such as purchases, payments and transfers, which data warehouses work really well with. In addition to this, data lakes can also easily become large data swamps which are more like a dumping ground with poor data integrity, poor quality, limited governance and metadata management, and inadequate data protection.

Too technical? Let's lean on an analogy as usual to understand the difference. Imagine a big library. The US Library of Congress or the British Library, perhaps. The entire structure is filled with rows and rows of neatly stacked books—ceiling-high, all properly segregated, indexed and organized in alphabetical order, categorized by topic or genre. The highly organized set-up makes it easy to identify and locate the book that you are interested in and you have guides to navigate you throughout the large expanse, to reach the spot where the book is stacked. That is synonymous with a data warehouse.

A data lake, on the other hand, is like a large natural lake or reservoir where numerous streams and rivers from different directions converge and the water accumulates. All kinds of streams, big and small, converge into the lake and become part of the lake itself. The different kinds of streams are not treated in any specific way depending on which stream they belong to. A stream could very well be from a tributary, melting glaciers, rapid or a waterfall—they all converge into the lake. And this lake becomes a common source for all kinds of activities like swimming, fishing, boating, irrigation, etc.

Owing to complementary benefits of both approaches, over time organizations started maintaining both these structures simultaneously and linking the systems together. This often led to data duplication, security challenges and additional infrastructure expenses. As a consequence, there emerged an opportunity to

amalgamate the best of both worlds, giving rise to **hybrid structures**.

A hybrid model is like a big supermarket that holds both packaged, branded products (structured data) and store-owned products like groceries, fruits, vegetables, grains, pulses, etc. (unstructured data). While the supermarket holds all kinds of packaged products, neatly organized and labelled in shelves in designated sections of the store, it also holds the unbranded products that the supermarket sells under its own brand name, stored in a less structured format, in bulk, to be weighed and packed later as per each customer's requirements. Customers, therefore, have an option to pick and choose the combination of branded and unbranded products from across various categories as per their requirement at the time of purchase.

Technology advancements over the years, especially the move towards cloud, has brought about a significant shift in the data-management process. Organizations can now adopt a hybrid approach that enables them to leverage the best of both worlds. The growing popularity of Databricks and Snowflakes is a testimony to the effectiveness of hybrid structures.

Databricks: The data lakehouse architecture

Databricks, a pioneer in data lakehouse, offers a way to leverage the scalability, flexibility and cost-effectiveness of data lakes while also addressing the governance, security and analytical capabilities traditionally associated with data warehouses.

It unifies data management by combining features of data lakes and warehouses, facilitating analysis of both structured and unstructured data sources. This is a cloud-based solution that provides scalability and flexibility, enabling organizations to adjust resources according to demand, ensuring consistent performance and cost-efficiency. The schema-on-read approach (where structure is applied to data only on retrieval from the stored location, rather than when it goes in) enhances agility in data exploration, allowing for raw data storage and on-the-fly schema application during

retrieval, reducing data preparation time. This architecture also lays the foundation for advanced analytics and machine learning by enabling the integration of structured and unstructured data.

Furthermore, it addresses data governance and security concerns through features like data lineage, access controls and auditing, supporting robust governance practices and ensuring data integrity and compliance. Apart from Databricks, the data lakehouse ecosystem of providers is growing with solutions like Google

BigQuery,^{*} Azure Synapse,[†] Snowflake and Amazon Redshift.[‡]

Snowflake: Cloud-native approach to data warehousing

Snowflake is a cloud-native data warehouse that takes a hybrid approach to data storage. It provides the scalability and elasticity of cloud storage offered by major cloud service providers like AWS, Azure and GCP. Additionally, it offers some of the key features typically associated with data lakes like real-time data analytics and support for semi-structured data formats. Snowflake's architecture separates storage and compute, which allows for independent scaling of each component. This means that organizations can scale up or down quickly based on workload demands, without affecting the underlying storage. This also enables data warehouse to support near real-time data analytics. Snowflake supports semi-structured data through its flexible data schema allowing data to be stored in a variety of formats, including JSON, XML and CSV.

Other notable players include Amazon Redshift, Google BigQuery, Microsoft Azure Synapse Analytics, IBM Db2 Warehouse on Cloud,^{*} and Oracle Autonomous Data Warehouse.[†]

5. Physical to logical integration

The rapidly growing variety of data sources brings along great opportunities for organizations, but as sources continue to multiply and become more diverse, integrating them becomes a greater challenge. So how do we address that? One of the key messages of

the book is the importance of approaching data initiatives using both logical and physical approaches rather than viewing it as a physical problem alone. The same concept applies here as well. Building a data stack by bringing all the data together in an operational data store for all use cases is not only a complex exercise, but can also become a never-ending process. Organizations are rapidly realizing that, whether through their own experiences or others'.

This is why taking a more logical approach to data integration is the need of the hour. In this approach, integration of data varies depending on the specific use cases or business requirements. Data mesh/fabric users say that they have achieved greater success in their integration efforts—78 per cent reported that they are successfully integrating their data platforms.⁵ I have talked about this decentralized data architecture in detail in Chapter 8, Multi-Source Data.

6. AI experimentation to AI first

In Chapter 3, Value Reimagined, I discussed the transformative potential of Gen AI for organizations and why Gen AI is the way to build an AI-first organization. This shift is driving organizations across industries to reassess their data stack, the fundamental infrastructure for effectively leveraging and managing data. To fully leverage the Gen AI opportunities, organizations have to ensure that their data infrastructure is capable of supporting it. For that, the data stack needs to have the following capabilities:

- **Enhanced data collection capabilities:** Foundational Gen AI models are trained on large open data from the internet, which gives it the remarkable ability to mimic human response anchored on vast amounts of information. But for business-value realization, these models likely need to be trained on enterprise-specific data. Hence, organizations need to identify and capture relevant data sources for Gen AI. This may involve

expanding data collection efforts to include diverse data types, such as text, images or audio.

- **Data processing and augmentation:** Data captured for training Gen AI models needs to be processed and augmented in a specific way to ensure the data is in a suitable format for Gen AI tasks. Modern data stack should be able to perform data augmentation techniques by applying transformations, introducing variations, or generating and leveraging synthetic data to increase the diversity of the training data for Gen AI models.
- **Computation resources:** Gen AI models often require significant computational resources, such as powerful CPUs or GPUs, to handle the computational complexity of training and inference. Distributed computing frameworks like Apache Spark or TensorFlow Distributed^{*} can be utilized to parallelize and distribute computations.
- **Model deployment and integration:** Modern data stack should enable seamless integration with Gen AI tools, including data storage, processing and analysis tools. Organizations need to ensure compatibility and interoperability between Gen AI models and other components of the data stack. Enterprises can leverage transfer learning and fine-tuning techniques to tailor Gen AI models to their specific needs, adapting pretrained models to their data sets and refining performance. This customization enables the creation of personalized Gen AI models that align with enterprise requirements and characteristics.
- **Continuous monitoring and iteration:** Organizations need to establish monitoring processes to track the performance and behaviour of Gen AI models in real-world scenarios. Modern data stack should facilitate the development of feedback loops, enable user testing and support continuous evaluation. These capabilities are crucial for identifying areas for improvement and gaining

insights to inform model iteration. Regular updates and retraining are necessary to adapt to changing data patterns or evolving user requirements.

- **Governance and security:** Organizations must establish clear guidelines and frameworks for the use of Gen AI in the data stack. This includes ensuring compliance with relevant data protection regulations, obtaining appropriate consent when using generative models with sensitive data and implementing safeguards to prevent unintended biases or discriminatory outputs. In my recent client conversations around Gen AI, while clients express great enthusiasm for this technology, a common concern that comes up in every conversation is on security, even on more internally oriented use cases.

7. Analyst teams to self-serve analytics

When all the layers of the data stack have been modernized, how can the consumption layer be any different? Gone are the days when the business owners were required to be dependent on the data experts or a specialized team to generate insights for decision-making. In the dynamic, fast-moving business environment, decision-makers at every level need to be empowered with data and equipped with the know-how to leverage it to make decisions and drive actions with speed and agility. And as a majority of business owners come from non-technical backgrounds, organizations are increasingly adopting self-serve analytics tools to enable the business owners and users to interact with the data as and when they need and generate insights as per the requirements of the use case. These self-serve analytics tools have contributed to the modern data stack by enabling users to seamlessly connect to various data sources, including databases, cloud services and data warehouses. Users can interact with and manipulate and shape the data according to their specific requirements without needing technical expertise. They can explore and visualize data through interactive dashboards, charts and graphs using built-in templates

and features, and perform ad hoc analysis in real time. I have explored this shift in detail in the upcoming Chapter 15, Data Democratization. So, stay tuned for more!

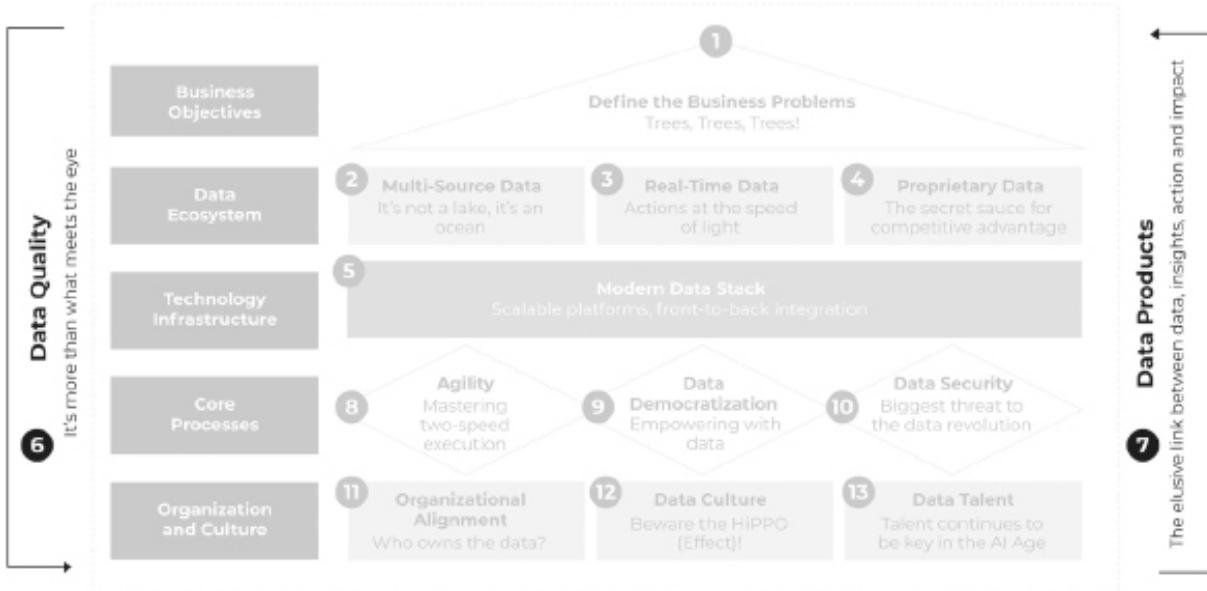
Key takeaways

- Traditional data stacks that are designed for narrow, structured data sets and working in a batch mode, often fall short in effectively handling the volume, variety and velocity of data, and are ineffective in enabling organizations to deliver transformational value.
- Modern data stack consists of loosely coupled yet tightly connected layers of the data management value chain, hosted on the cloud, which enables rapid scalability, accessibility and adaptability.
- There are seven key technology shifts that an organization must make to build a modern data stack:
 1. Harness the power of the cloud and cloud-enabled tools, technologies and services made available by the hyperscalers, to build asset-light data infrastructure that is easily scalable and elastic.
 2. Shift to real-time data infrastructure to generate insights and drive decision-making and actions in real time.
 3. Moving from a monolithic to microservices architecture to decouple different layers of the stack, allowing organizations to quickly adapt to changing business needs.
 4. Adopt hybrid data structures like the data lakehouse approach to combine the strengths of data warehousing and data lake architectures, facilitating better storage and integration, leading to effective utilization of data across the organization.

5. Adopt a logical approach to data integration, depending on the specific use cases or business requirements, to be more effective with data.
6. To leverage the full potential of Gen AI, organizations must ready their data infrastructure, with enhanced data-management capabilities, compatibility and interoperability with Gen AI models and greater focus on security.
7. Enable self-service tools for business users to seamlessly interact with the data and extract insights specific to the business requirements with speed and accuracy.

CUTTING ACROSS LAYERS

DATA QUALITY AND DATA PRODUCTS



'Integrators across the solution layers'

12

Data Quality

It's More than What Meets the Eye

'More data beats clever algorithms, but better data beats more data.'

—Peter Norvig,
Research director at Google

The era of Big Data brings with it unlimited opportunities for organizations to derive deep insights to make data-driven decisions. But it also brings with it the complexity of dealing with huge and diverse data sets being generated and used on a continuous basis. Copious amounts of data that is being fed into various systems and tools needs to be of high quality through and through. Faltering at any stage would result in a ripple effect trickling down the entire data value chain, undermining the foundation on which the organizations base their decision-making, both strategic and operational. As a result, the role of data quality, which has been a long-standing issue, does not diminish but becomes even bigger in the Big Data world.

How? Let's dive right in and find out, shall we?

The promise of Big Data

A fundamental concept of statistics called 'the law of large numbers' which describes the behaviour of averages of numbers, states that as the sample size (n) increases, the average of the observed values will converge to the expected value of that variable. In simple words, the bigger the sample size, the more representative it is of the population (N), and the better the chances are to get closer to a

reliable estimate. And based on this, inferences and assumptions are made about the population, helping us understand any trend in the long run.

With the advent of Big Data, the size of these sample populations became bigger and bigger. Now that we had data in troves, Big Data brought about a scintillating promise that with larger sample sizes, the insights generated would be more accurate and reliable. With a large enough sample, the estimates derived are more likely to reflect the true characteristics of the population providing organizations and researchers more confidence in their findings and thus making more informed decisions.

Agreed, it does sound logical. And organizations have been leveraging Big Data, investing in tools and technologies to analyse all this data at scale, generate insights and deliver them to the users through interactive visualization tools. But if you dig deeper, fewer are those who claim that these insights generated are truly 'actionable' or align to the outcome they wanted to achieve.

While Big Data has enabled organizations to successfully leverage insights for directional purposes, it has not fully lived up to its expectations when it comes to translating them into specific actions.

Why has Big Data failed to deliver on the promise?

While the world has been rejoicing at the abundant availability of data that can very well be their ticket to more accurate, untapped or unique insights, as discussed in Chapter 11, Modern Data Stack, both the nature of data and business expectations have completely changed. And that is where the universally accepted theory of law of large numbers has fallen short in solving the data-quality issues:

Complexity brought on by variety and granularity

With the explosion of data, integrating and harmonizing highly diverse data, validating the accuracy and completeness of each data point, and cleaning such diverse data sets has become a big challenge. Maintaining data quality across the data-management

value chain has become an increasingly complex and tedious task. For example, take the millions of data points an e-commerce company deals with every day—their inventory data, customers browsing activity, purchase history, ratings, reviews, social media feeds, shipping data, logistics data, etc. And within each are a number of subcategories and sub-subcategories of data that add to the complexity and granularity. Imagine how colossal is the task of integrating this data and translating these into actionable insights, considering Amazon collects twenty-three different types of data points on each of its customers on a continuous basis.¹ Multiply this with the 300 million individual customers that Amazon has today (As per 2022 figures).²

'Big' Data is expected to drive 'individual' actions

For many years in the realm of science and research, particularly in scenarios where data was limited (often referred to as the 'small data' era), the problem-solving method was primarily on the proving or disproving hypothesis. This process heavily relied on establishing causality to understand why certain events occurred and to build a 'cause-effect model'. However, the Big Data era brought about a significant shift due to the sheer abundance of data available. With such extensive data availability, it became possible to uncover deeper and more robust correlations, often negating the need to delve into the intricacies of causality. Organizations and researchers could now confidently make decisions and draw insights, make predictions and make real-time decisions based on these strong and meaningful correlations.

However, while Big Data analytics has opened doors for organizations to benefit from new, directional insights based on correlations, when it comes to translating these insights to specific actions, especially when actions need to be driven at an individual level, the burden of proof—the obligation to provide sufficient evidence or justification to support a claim or decision, is higher. Therefore, generic, or broader level insights are not effective here.

Let me give you an infamous faux pas in Google's history of leveraging Big Data insights to drive action at the individual level.

In 2008, Google launched an ambitious project called Google Flu Trends (GFT), to predict flu outbreaks using search patterns and correlating them with actual flu activity. It aimed to offer real-time insights for faster public-health responses. However, during the 2012–13 flu season, it significantly overestimated flu activity, by as much as 140 per cent off the mark, compared to the actual data from the US Centers for Disease Control and Prevention (CDC). This is because of some inherent flaws in its algorithm that was not trained to distinguish between data from search for actual illness and those driven by curiosity or media influence. This failure underscored the importance of considering situational factors when developing predictive models. It also highlighted the inherent challenges of relying on passive data collection and the need for a more robust data-quality process.³

But while GFT wasn't a roaring success, it paved a way for others to leverage the model to build more elaborate models on top of that. In fact, in 2013, a team of medical and data science researchers succeeded in building a better influenza model by adding more variables to GFT, like meteorological data, temporal variables like seasonality, etc. According to them, GFT was the sole external source of information that held statistical significance. GFT's primary purpose was to offer additional signals or insights that complemented existing data, which it did quite well.⁴

Therefore, when translating directional insights into individual actions, the burden of proof is higher and it is important to triangulate these insights by combining data points from multiple sources to be more comprehensive. It adds to the robustness of the model and enhances the confidence in the insights drawn.

In a world where we are rapidly moving towards a 'segment of one', every individual is unique and therefore the products and services are increasingly being customized for each individual (which I talk about in detail in Chapter 21 of Section III, The World of

Hyper-Personalization). It is therefore critical that the generated insights are robust enough to enable individual actions.

While we started with a word of caution on using Big Data, clearly there are many successful implementations to drive from insights to individual actions. Amazon's use of Big Data in its recommendation engine is a fairly successful example of leveraging Big Data to drive individual actions. Amazon churns millions of data points on its customers to drive its recommendation engine. But the trick lies not in how much data Amazon has, but how it goes about using this data intelligently. Instead of solely relying on each customer's purchase history alone, Amazon focuses on understanding each customer by building their comprehensive profile. The belief is that if Amazon knows who its customer is it can better predict what they are likely to buy. So, a 360-degree customer view is built by analysing multiple data points on each customer, like the customer's purchase history, browsing activities and combining them with census data sets like demographic details, shipping address, etc. In addition to that, using 'collaborative filtering', Amazon matches similar customer profiles to identify common purchase preferences, to suggest relevant products. Such rigorous methodology, based on strong and robust correlations supported by multiple factors, enables Amazon to successfully drive recommendations for each individual. This is why it can claim that as much as 35 per cent of its revenues comes from its recommendation engines, either online or via email.⁵

We see from the Amazon example that to drive actions at an individual level, data needs to be at a granular level and take a multivariate or a 360-degree view. Additionally, for effective translation of data to insights or actions, the data quality needs to remain intact across the entire data value chain. Therefore, given the opportunity to drive individual actions from Big Data, the very scope of data quality has expanded and become highly complex.

Reasons behind data-quality issues

The quality of data can be compromised anywhere across the data-management cycle. It can be caused due to human error, anomalies, malfunctions or even data overload. The various stages and the ways data quality can be affected are:

At source

Quality issues can arise right at the start, from the point of origin or the source of data. The various ways data can be vulnerable are:

1. **System malfunction or downtime:** The machine or device that is being used to capture data can very well malfunction or face downtime due to unplanned stoppage. At this time, the device would fail to collect important data leading to missing or inadequate data.
2. **Bad source:** Quality issues can also arise if the source from which the organization is collecting data itself does not adhere to data-quality standards, leading to inherent quality issues.
3. **Machine errors:** To digitize large amounts of data quickly, organizations depend on OCR (optical character recognition) technology, to scan images and extract text from them automatically. In such cases there are obvious chances of translation or recognition errors due to misinterpretation or misattribution.
4. **Ambiguity issues:** While newer attributes are added every day, owing to the rapid expansion of the length and breadth of data, certain attributes may not be captured at all, either by humans or machines, because these have not been defined yet.

During processing

Once the data is captured, the way it is handled or processed within the organization is also critical to ensure quality. There are multiple ways data quality can be impacted:

1. **Data-entry errors:** Data-entry errors may arise as a result of human mistakes, including typographical errors, inaccurate data input and improper data formatting. Data-entry errors can additionally manifest as a result of machine-related issues, like inaccuracies during data imports.
2. **Data-transformation error:** Data-transformation errors materialize when data undergoes conversion from one format to another. These are errors that arise due to inaccurate data mappings, improper data conversions or erroneous data transformation rules.
3. **Lack of data literacy:** The person who is in charge of handling the data may not be adequately trained to work with data as they don't understand what certain attributes mean.
4. **Obsolete data:** In the fast-moving world, data ages rapidly. It is critical that the data be collected and processed as soon as it is generated. Any delay in capturing or updating the data may lead to value erosion or redundancy.
5. **Data overload:** Like I have said multiple times in this book, while abundance of data is a boon, too much data is a bane. Both humans and machines can be overwhelmed by the scope of data they are required to deal with resulting in lapses or errors
6. **Deviations and anomalies:** At times deviations and anomalies in data processing can occur due to various reasons like programming errors, data inconsistencies, unexpected or unrecognized data formats, or issues in data integration. This can lead to results that are

inconsistent with what is anticipated, expected or considered normal.

During cross-system integration

Another stage where data quality is highly vulnerable is when it flows from one system to another. The multiple ways data can be compromised at this stage are:

1. **Data loss or leakage:** The most common issue is the loss of data while integrating data from various systems. This can happen due to accidental omission, data filtering or exclusion rules, or technical issues during the migration process.
2. **Format inconsistency:** When you are dealing with multiple sources of data, it is common to find differences in the same information between those sources. These differences can include variations in formats, units of measurement, or even spellings of words, which can cause data-quality issues.
3. **System-compatibility issues:** When multiple systems are used to transfer and integrate data, it is not necessary that every system is compatible to the formats used by the data source, causing issues while integrating these.
4. **Inadequate treatment:** While consolidating data from multiple sources, it is critical that clear standards or rules are set to treat duplicates or missing data. Failure to do so can create inconsistencies and errors.
5. **Aggregation error:** While collating data from multiple data sets, several fields may be generated or calculated based on other fields within or from across systems. These formulas are executed automatically whenever new data is entered or updated in the related fields. But if there is any error in these formulae, it can result in the entire data set getting compromised.

It is therefore critical to monitor the data quality, not just at the source, but during the entire lifecycle of the data, because it has the potential to have a ripple effect down to the decisions made or actions taken based on it. So let me now get into the way data quality can be effectively managed.

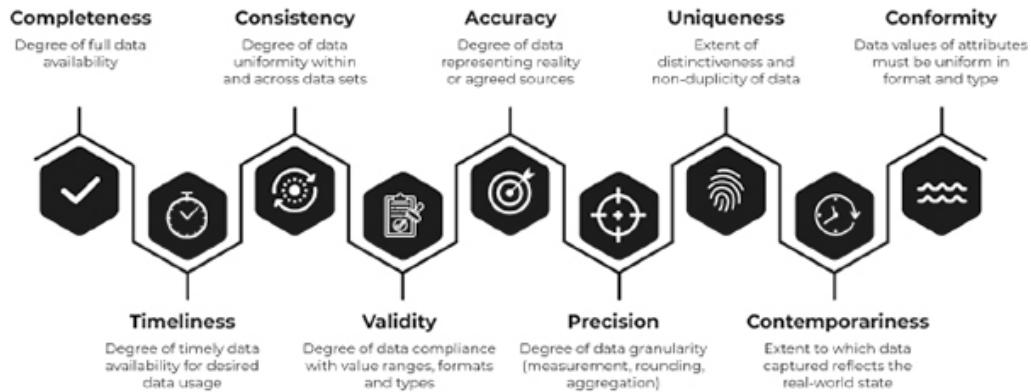
The traditional dimensions of quality still hold true

In the world of Big Data where organizations heavily rely on data-driven decision-making processes, it is critical to ensure that the underlying data that is being fed into the analytical models strictly meet the quality benchmarks. Therefore, monitoring and ensuring the quality of data becomes critical at every stage of the data management cycle.

Now, traditionally the quality of data has been evaluated on nine key parameters. Let me briefly take you through these dimensions:

1. **Completeness:** As the name suggests, completeness, one of the frequently used dimensions of data quality, evaluates the extent of missing data. Of course, you might not have 100 per cent of the data, at all times, but the missing data should not impact the analysis through bias or by omission. For example, in a census data, let's say the population of a state is 1,00,000 and the complete demographics are available for 99,486 citizens. So, the percentage of completeness in this case is 99.48 per cent, which is pretty good to derive useful insights from these.

Traditionally, data quality has been defined on nine dimensions



- 2. Timeliness:** Timeliness measures the lag between when the data is generated vs when it is made available for use—a measure to avoid old data. For example, the movement of share price needs to be made available to the traders in real time as they need to make timely decisions when buying or selling shares. Real-time share price data allows them to assess the current market conditions, track price fluctuations and determine the best entry or exit points for their trades. In case the information is delayed, it could result in huge losses for the investors.
- 3. Consistency:** It refers to the uniformity and coherence of data across different sources, systems and timeframes. It ensures that the values, formats and types of data are aligned and harmonized to ensure accuracy and reliability. For example, a retailer who maintains data on purchases made by its customers both online and offline, should ensure that the customer information, transactions history, loyalty details, etc., captured within the system are accurate and reliable.
- 4. Validity:** It measures how well data meets certain criteria. Validity of the data can be evaluated against established guidelines, such as predefined limits, the order in which events occur, or rules set by the business. In simple terms, it's about making sure the data fits the expected patterns and rules, identifying or eliminating

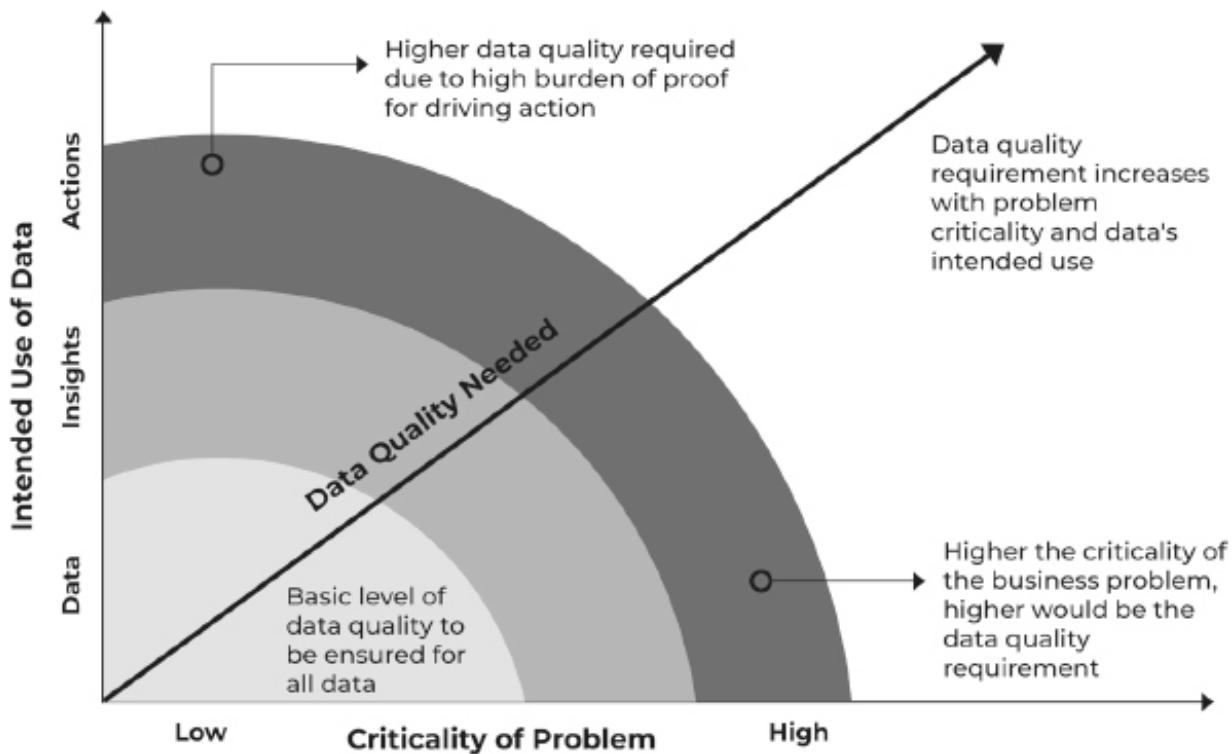
the outliers. For example, the demographic data on customers of a nightclub cannot be less than twenty-one years old, or more than 100 years old (very unlikely). Or an order is shipped after it is placed. So, a shipping date prior to the order data becomes invalid.

5. **Accuracy:** Accurate data refers to error-free records that represent the true and exact values, attributes or characteristics it is intended to capture, without discrepancies or distortions—the measure to avoid wrong data. For example, accuracy of an address is critical for e-commerce companies to make deliveries to the right customer.
6. **Precision:** Precision measures the degree to which the data has been rounded off or aggregated. For example, data on financial transactions must be precise to the second to monitor activities in real time. One of the many reasons for which it requires such precision is to raise alarms against fraudulent transactions in real-time.
7. **Uniqueness:** This dimension ensures that each data record or entity is distinct and free from duplication, to avoid double counting or misreporting. For example, if a customer data gets recorded twice, once in name-surname format and the second time with surname-name format, it would still be the information on the same customer but may get counted twice in analysis.
8. **Contemporariness:** It ensures that the time of data collection corresponds accurately with the time of it being created or recorded. For example, the stock market prices should have the time and date stamp of when the data was generated to ensure its recency.
9. **Conformity:** It ensures that data of the same attributes should be stored in uniform format and type. For example, if the dates are stored in MM/DD/YYYY first, then that is the format which needs to be followed throughout.

While these dimensions stay relevant and critical to ensure data quality, there is a crucial aspect that we are missing here. The growing complexity of incoming data and the failure to translate generic or broader level insights into individual actions, requires the purview of quality to become bigger and more fluid. And according to me, there is an overarching dimension which should be the starting point that would ensure data quality throughout—the ‘context’ or the outcome that the business is trying to achieve. Once that starting point has been established, the rest of the traditional dimensions can flow from there.

Data quality requires a context-first approach

Context should drive Data Quality to ensure that data is ‘Fit for Purpose’

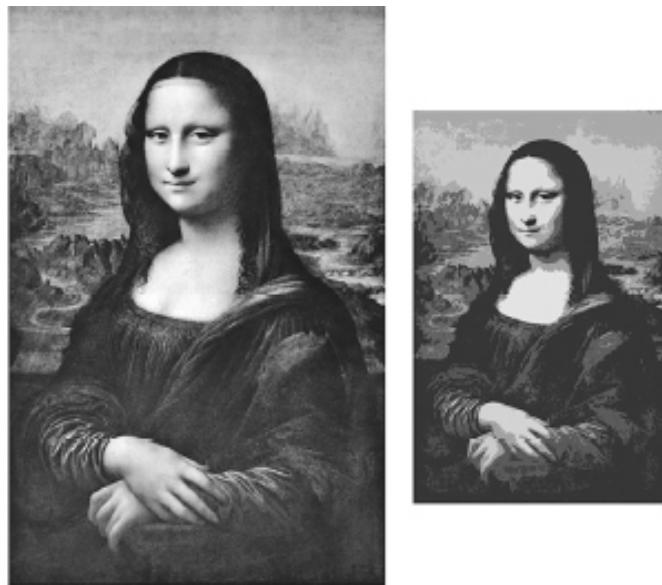


Data quality can no longer be measured in absolute terms. Context plays a crucial role in determining the relevance and appropriateness of data for a specific purpose or use case. In other words, data quality should be ‘fit for purpose’. What may be considered high-

quality data in one context might not hold true for another. It is driven by the circumstances in which the data is being used.

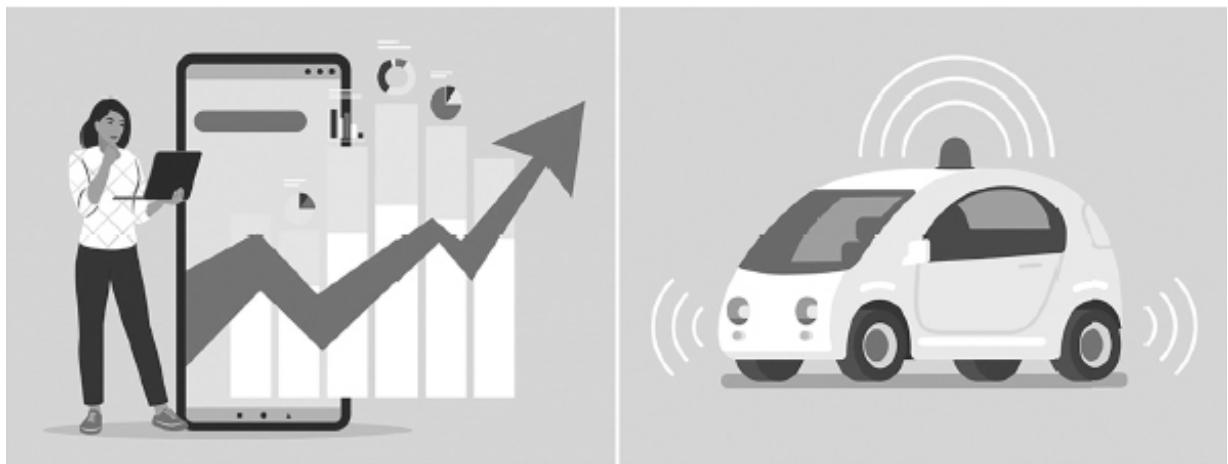
When determining the context of data quality, organizations must consider two important factors:

1. **Intended use of data:** One important factor is the intended use of that data, that is, what is the purpose of that data or outcome expected. And based on that, the quality of data is evaluated. A fitting example is choice of resolution when sharing an image for further use. The purpose of that image would determine the ideal resolution. If the image is expected to be printed as a banner, it needs to be shared in the highest resolution. But for use in a PowerPoint presentation, lower resolution would be preferred owing to lower file size that is easier to manage and use.



2. **Criticality of the business problem:** The other important factor is the criticality for the business—to what extent is the quality critical for the outcome or intended use. For example, in cases where data is being used for self-driving cars, it is important that every datapoint that is being captured and processed should be

of the highest quality in terms of most of the nine dimensions that I talked about above. Because in such use cases, both the criticality of the problem as well as the burden of proof is very high. On the other hand, if we need data on customers to run a marketing campaign, it will not require the same level of accuracy or precision or timeliness.



Context determines the dimensions to focus more on

The weightage of each dimension also varies based on the context in which the data is being used. For example, in supply-chain management, data quality would include evaluating the accuracy, completeness and consistency of inventory data, shipment records and supplier information. In contrast, while monitoring credit-card transactions, timeliness becomes a bigger dimension in addition to accuracy and consistency of data.

Starting with the end in mind

As I have emphasized multiple times in the book, to navigate well in the Big Data world, one has to start with the end or the outcome in mind. The push and pull between wanting to capture as much data as possible vs ensuring the data being utilized is always of the highest quality requires a fine balance. So, starting by establishing

the context in which the data is expected to be used can help organizations tailor their data quality efforts to the specific needs and business objectives. This not only ensures that the data meets the relevant standards and requirements at all times, but also helps organizations significantly narrow down their data efforts, saving both time and resources.

Data quality—necessary for organizations becoming AI-first

In the Big Data world, where AI and machine learning continue to rise in prominence, data quality has become ever more critical for organizations that are aiming to become AI-first. Over the past few years, the number of organizations adopting AI has grown significantly, 2.5 times higher than in 2017,⁶ with organizations using AI in at least one business area. It is evident that organizations are increasingly seeing the benefit in leveraging and scaling AI across their value chain. Another recent survey revealed that 80 per cent of executives surveyed thought that automation could be applied to any business decision, reinforcing the potential of implementing AI at scale in any organization.⁷

However, the effectiveness of an AI model depends on the quality of input data. In the past, AI adoption has often been dictated by data availability rather than outcomes. Data quality is a major challenge hindering scalability of AI, even more so than how it impacts analytics. If data is the foundation of the AI age, data quality is clearly one of the biggest bottlenecks. Another reason the topic of data quality is so critical as we enter the AI age.

But AI can also be a solution for Data Quality!

Although we have seen multiple evidence of how AI can solve various business problems effectively, there is one use case that has been less ‘glamorous’ but extremely important—the growing role of AI/ML in improving data quality. By using AI- and machine learning-

enabled platforms, organizations can significantly improve the process of managing data quality across the nine dimensions and through the various stages of the data-management cycle. A recent survey of 1900 data practitioners and C-level executives revealed that almost half 48 per cent of the companies surveyed are using data analysis, ML or AI tools to address data-quality issues.⁸

Following are some prominent ways in which AI and machine learning can help improve and automate the process of managing data quality:

1. **Data cleaning and processing:** AI/ML algorithms can be used to automate data cleaning and processing tasks. They can identify and handle missing values, remove duplicates, outliers and inconsistencies in the data. They can also automate data transformation such as standardize and normalize data.
2. **Anomaly detection:** AI/ML solutions can help identify anomalies or outliers in the data. By establishing patterns and detecting deviations from those patterns, AI can help identify errors, fraud or unusual events in the data set. Take PayPal, the digital payments company. It leverages AI-based anomaly-detection techniques to analyse transaction data, user behaviour patterns and historical data to identify anomalies, such as unusual purchasing patterns or suspicious account activities. This enables the company to proactively flag potential fraud instances, protecting both the company and its users. In fact, 83 per cent of respondents recognize ML as pivotal to their company's e-commerce fraud strategy, in a survey conducted by PayPal and Forrester.⁹
3. **Data validation:** AI/ML techniques can be employed to validate the accuracy and consistency of data. By training models on test data or established rules, it can help identify discrepancies or errors in the data set. For example, for an e-commerce company, an ML solution

can check the consistency of the addresses, flagging any entries that do not conform to the expected patterns.

4. **Data augmentation:** Gen AI is being employed to generate synthetic data that can be used for testing and validation purposes without risking the exposure of sensitive or confidential information. It is estimated that by 2024, 60 per cent of the data used for the development of AI and analytics projects will be synthetically generated.¹⁰ For example, Waymo, a leading autonomous driving technology company, uses AI-generated synthetic data to simulate various extreme scenarios around road conditions, weather conditions, complex traffic scenarios, rare events and other potential safety risks that are difficult to replicate in the real world.¹¹

5. **Natural language processing (NLP):** AI/ML models based on NLP techniques can improve the quality of data by performing tasks like sentiment analysis, text classification, etc. For example, an NLP solution can analyse customer feedback and complaints to identify trends and sentiment, helping companies improve their services and address quality issues.

6. **Data governance:** AI/ML solutions can assist in managing metadata. They can automatically classify and tag data, identify sensitive information and help enforce data-governance policies. AI/ML models can be deployed to continuously monitor data quality in real-time.

Using AI and ML can enable organizations to create a virtual cycle of managing data quality where the systems not only continuously monitor the quality but also learn from it to continuously improve and enhance their processes.

Granted, AI alone cannot be the be-all and end-all to improving data quality. It still needs to be used in conjunction with traditional methods like combining it with domain expertise and knowledge of

data professionals and establishing robust frameworks, policies and procedures to ensure data quality. For example, AI may be very good at identifying data patterns or performing complex analysis but might not be effective with the context. Here the traditional role of business or domain experts would add value by providing deeper understanding of the data elements required. While AI algorithms can automate certain aspects of data cleansing, traditional methods, such as manual review and correction, are still required in many cases, when handling complex data transformation tasks or making subjective decisions when dealing with quality issues. Furthermore, human expertise and domain knowledge can also help refine these algorithms continuously as the business requirements and industry dynamics continue to evolve. This hybrid approach of integrating AI and ML with traditional methods can dramatically reduce the efforts and improve the process of managing data quality.

Key takeaways

- Big Data offers opportunities for insights, but its complexity requires a heightened focus on data quality. The role of data quality in the Big Data world does not diminish, but becomes even bigger, and cuts across the entire data-management cycle.
- In the data-first world, the concept of data quality goes beyond traditional dimensions. It becomes more fluid and dependent on the specific context and desired outcomes.
- Starting with the desired outcome in mind helps organizations prioritize data quality based on requirements. This context-first approach helps identify key quality dimensions for prioritization.
- Adoption of AI and ML technologies can enable organizations to effectively manage data quality. These technologies cannot just monitor data quality but help continuously improve quality by creating a feedback loop for ongoing enhancements.

13

Data Products

The Elusive Link between Data, Action and Impact

'A great product isn't just a collection of features. It's how it all works together.'

—Tim Cook,
CEO, Apple

Managing data quality is one of the biggest challenges of the Big Data world, to which I dedicated the previous chapter.

In this chapter, I am going to talk about what I believe can become one of the most effective solutions to maximizing value from data in the Big Data world. The answer lies in adopting a systematic approach to establish a clear link between the various layers of the data stack, anchored in business outcomes, to expedite the data-management cycle. And the most effective way to do so is through 'data products'.

Productization of data

What is a product? In simple terms, it is something tangible or intangible that is created, designed or manufactured to fulfil a specific need, want or demand. It can be a physical object, such as a car, smartphone, or piece of furniture, or a digital product like software applications, eBooks, music downloads, etc. What distinguishes a product is its ability to provide a standardized and repeatable solution. This standardization ensures that the product can be consistently reproduced and applied, making it reliable in fulfilling its intended purpose each time.

Now who could have ever imagined that data—once considered a mere input or byproduct of business decision-making—would emerge as a vital resource. While the use of data started as a means to solving a business problem, its proliferation across every aspect of business has made it an essential ingredient for decision-making and driving action. And as more and more data was utilized, there emerged multiple use cases where the same or similar sets of data elements were being used. It could now be collected, processed and utilized repeatedly to address a multitude of challenges and opportunities.

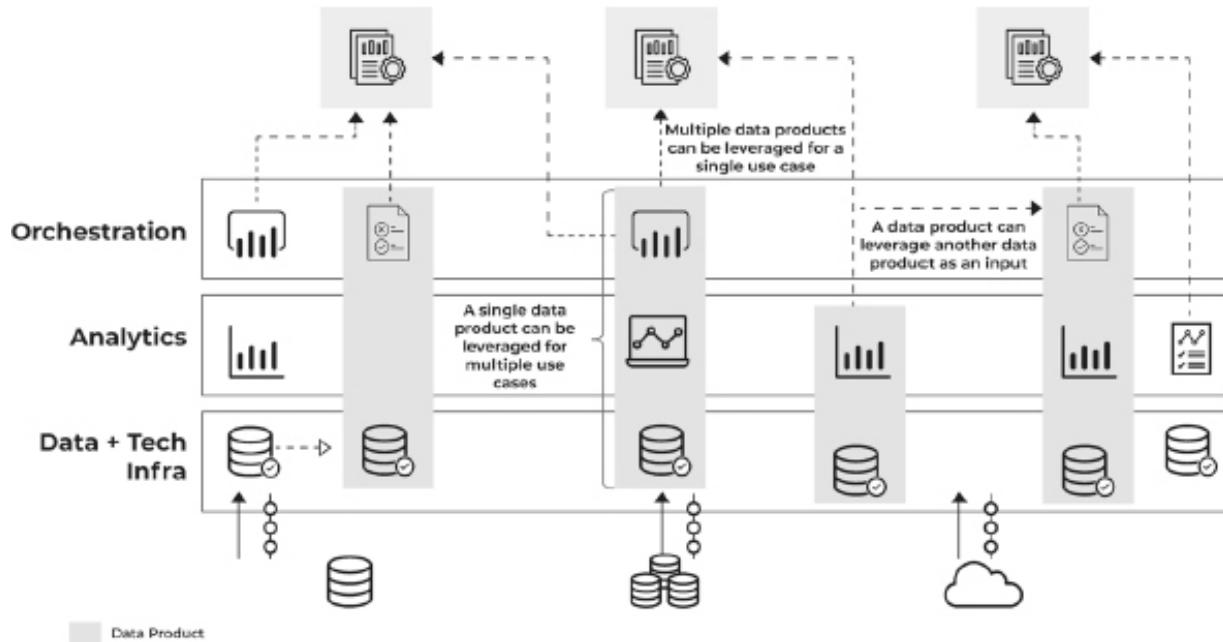
Recognizing such repeatable use cases and applications of data, organizations started seeing value in the productization of data, to enhance its reusability and drive efficiency through the process. Whether it's for optimizing supply chains, improving customer experiences or enhancing predictive analytics, data's repeatability makes it an ideal candidate for productization.

While the concept of data products is relatively new, it has been gaining momentum rapidly. Today, data products are no longer a novel idea but an integral part of business strategies, empowering organizations to unlock the latent potential of their data assets.

Let's define data products

Data products are digital assets built as a vertical slice, integrating various elements across the data stack to deliver business use cases in a repeatable manner and accelerate data to insights to actions to impact cycle.

What is a Data Product?



In simple words, data products contain data packaged in a more structured way to solve business needs and can be reused multiple times. Data products are a way to transform data into pre-built solutions. These are built by vertically combining multiple components of the data stack and can be used multiple times, for multiple use cases, making it scalable.

Let's look at some salient characteristics of data products that make them such a valuable construct:

- 1. Outcome-focused:** The primary objective of a data product is to use data to facilitate an end goal. So, data products have a well-defined, well-articulated objective or goal. For example, the primary objective of a customer-recommendation data product is to use data on customers to facilitate their buying journey.
- 2. Actionability:** The data product goes beyond just presenting data, and focuses on delivering insights that drive specific actions, strategies or outcomes. For example, Google Analytics not only captures data on

users but translates it into insights based on user queries.

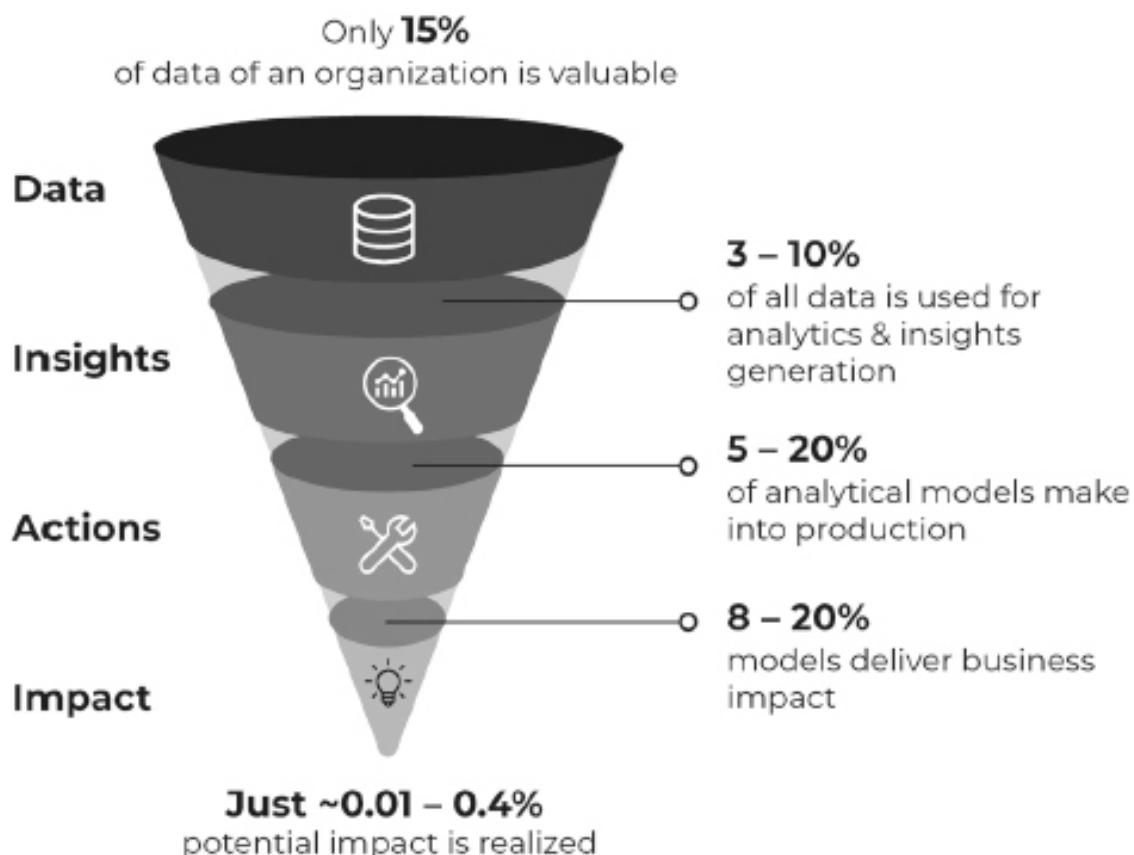
3. **Reusability:** Data products are packaged in such a way that they reduce manual effort, whether it is cleaning, processing or insight generation. These products can be leveraged multiple times for multiple use cases for instant insight generation or driving action. For example, a data product for customer segmentation can be used by multiple teams or for multiple product categories.
4. **Accessibility:** Since these data products are pre-built, well-packaged end-to-end solutions, they can be used by both technical and non-technical users, reducing external dependencies. For example, self-serve tools can be used by users across the organization to interact with data as per their business requirements. And it is critical to keep security and privacy in mind while doing so, which I will elaborate further in Chapter 16, Data Security.
5. **Integration and interoperability:** Data products are built in such a way that they can be easily integrated with any solution or can be combined with other data products to build a bigger solution. For example, a customer-segmentation module can be leveraged for marketing campaigns, customer retention, product offerings, etc.
6. **Measurability:** Another important aspect is the impact or performance of data products. There should be well-defined measurable KPIs for the data products that align with the goal or objective for which it is built. For example, if the data product aims to improve customer retention, KPIs would include customer churn rate, customer lifetime value or repeat purchase rate, etc.
7. **Iterative improvement:** Data products should continuously evolve through a cyclical process of feedback, learning and refinement. They should be iterative to ensure continuous improvement and

adaptation, to enable organizations to align better with the changing business and user needs.

Why data products?

As I have highlighted in Section I, organizations are investing huge amounts of time and money in figuring out the best way to realize the full potential of data, but most of them have failed to do so effectively. The translation of data to insights to actions and eventually to impact is extremely low for organizations leaving significant value on the table. This value can be realized by improving the efficiency not just within the layers but also by tightening the connection across the layers of the funnel. Let's first understand the value leakage across the layers of DIAI funnel.

Data to Insights to Actions to Impact (DIAI) 'Conversion Funnel'



To begin with, not all data an organization possesses is relevant for business. Estimates indicate that on average, less than 15 per cent of all the data collected by an organization is valuable for business.¹ As we move down the funnel, there is significant loss at each stage of the funnel. To begin with, estimates suggest that of all the data in an organization, merely 3–6 per cent is used for insight generation or analytics.² That is a substantial loss right from the word go. Even the 15 per cent of relevant data and 3–10 per cent of that data being used for analytics are numbers that are likely to change and perhaps further go down.³ This is because the nature of data continues to evolve, and more and more unstructured data is being added each day. 80–90 per cent of the data in organizations is typically unstructured now.⁴ Of that, less than 1 per cent is being used or analysed at all, unlike structured data where ~50 per cent is utilized.⁵ As the proportion of unstructured data will continue to rise, the conversion metrics at the top of the funnel will further diminish.

Moving down the funnel, as mentioned in the book earlier in Chapter 3, Value Reimagined, despite a large number of models being developed, only a few get implemented into the production systems. In fact, estimates suggest that somewhere between just 5 per cent to 20 per cent of all the models developed go into production.⁶ In my experience over the last twenty-five years, I have consistently observed that only a fraction of the models developed in sophisticated Fortune 500 enterprises (both organizations I have led and clients I serve) go in production. I believe, on average, seven out of every eight models developed are abandoned or wasted. The graveyard of abandoned dashboards, metrics and models is evidence that the link from data to insights to actions to impact is not very effective.

To make matters worse, the impact from the models that do make into production is typically much less compared to what was estimated in test environments. A massive 80–92 per cent fail to generate any value for the organization.⁷ On average only one out of

every ten model delivers any kind of business value. Now if we cumulate the value loss at every stage of the DIAI funnel, it is evident that the overall impact is **at best less than 0.5 per cent of the data available, but in most cases, it is tending to a measly 0.05 per cent, sometime falling to as low as 0.01 per cent**. Moreover, this impact realized is not from the total data that an organization has access to, but from just the proportion of data, typically 15 per cent, which is considered useful for business,⁸ which makes this low conversion number even more stark.

It is clear that despite the huge investment made in technology, the conversion from one layer of the data funnel to another remains sub-optimal. Now why is this the case, despite so much data and the best of technologies available to organizations today?

This is because often organizations focus on solving for a problem at one level of the funnel while missing out on the front-to-back integration across the funnel. For example, trying to bring all the data together at one place, which can become an all-consuming, never-ending task. In such cases, disproportionate focus is on getting their hands on all the data, while the business use cases and what data is required for them might not be well defined. This can create a gap between the data infrastructure you have built and what is needed.

And even if organizations are somehow able to collect all the data they need and do the required analyses, they are faced with the tedious task of figuring out the most effective way of integrating and delivering these insights to the right systems to enable decision-making and actions.

Improving the conversion of data across the value funnel is a huge opportunity. It can be done both by increasing the efficiencies at a particular stage of the funnel and better integration across the funnel. Conversion metrics are currently so poor that even incremental improvements within or across layers would result in significant uptake in the value realized.

Furthermore, while increasing the conversion metrics is necessary, it is important to realize the limitations. For example, one cannot

increase the proportion of data being used, right from 3 per cent to 50 per cent. Therefore, the efforts must be targeted. Instead of spending resources on bringing all the data together, converting all of it into models, putting all models into production and so on, focusing on the few most business critical problems, would give you much greater chances of success. Which brings me back to the importance of clearly defining the business problem, talked about in Chapter 7, Define the Business Problems, to narrow down the critical root cause, which would help in identifying the most critical or the most 'useful' data required to create maximum impact.

It is critical for organizations to figure out a way to plug this enormous value leakage, which requires various layers of the DIAI funnel to be well connected and to start with a clear outcome in mind. And in my experience the most effective way of doing that is by creating data products.

Data products are the most effective way to bring order to all this chaos

Data products are designed with the end in mind. Each data product helps achieve some goal or outcome. The outcome may vary, such as customer segmentation or product recommendation, but with clearly identified and defined outcomes, a data product is built backwards to identify the technology needed, the analysis required and the data to support that. The data product is then packaged as a self-contained unit.

Now since the data product is designed using this systematic approach, the connection between data, insights and actions is clearer and much tighter, which naturally increases the chances of creating the desired impact. In other words, data products significantly compress the various stages of DIAI framework, by leveraging capabilities across the horizontal expanse of the data stack into reusable vertical slices.

Data products are also a potential solution to ownership and collaboration issues

In the era of Big Data, ownership and collaboration is not easy to establish. With the growing complexity of the business ecosystem, multiple stakeholders are now involved in generating and utilizing data throughout the organization's value chain. As a result, a more structured framework around accountability and ownership is required to facilitate collaboration. Data products will likely emerge as a potential solution to address this ownership and collaboration challenge, offering a structured framework to clear accountability and responsibility for data initiatives. This approach would help avoid data silos, duplication of efforts and confusion about who is responsible for data-related tasks.

Moreover, with defined roles and responsibilities, cross-functional teams can collaborate seamlessly. Shared access to data, collaboration tools and workflows will enable seamless collaboration among team members, encouraging collaborative problem-solving, iterative development and collective decision-making. Stay tuned for more on this topic in Chapter 17, Organizational Alignment.

The different types of data products

There can be multiple combinations made by vertically combining the various layers of the data stack. Across the spectrum, data products can be broadly categorized into five groups, namely processed data, derived data, algorithms, decision support and automated decision-making.

While the first three are more suitable for a technical user, the latter two are more focused on business outcomes and are apt for both technical and non-technical users.

Now as we are entering the data-first world, most organizations aim to build data products that can compress the entire data-management cycle and are easier to understand and use by non-technical business users as well. These are ideally the data products

that either support decision-making or automate action. So, let me first talk about these two in detail.

Data-enhanced products for automated decision-making or action

These data products are the most advanced types of data products that trigger recommendations or automated actions. They leverage capabilities across the data stack, including data, technology, analytics, orchestration and consumption layers. These data products have an AI/ML driven core that does all the work and presents the user with the final output or the best immediate actions to be carried out.

They are designed with user experience in mind, allowing the user to easily interact with the system without dependencies on data or tech experts. These can also be enabled for automated action which does not require any user involvement. One good example here is a self-driving system in an autonomous car. And I am going to explain how one of the many data products that go into operating a self-driving car works. It is called the real-time control and actuation data product—how the actions of the car are controlled and executed in real time. It works in conjunction with other modules to execute actions based on the analysed sensor data and planned driving strategies. The sensor data about nearby vehicles, pedestrians, road conditions and other relevant objects is collected and processed using AI and ML modules. This data is analysed in conjunction with maps and GPS by another module to make decisions about what the car should do. Another one takes those decisions and figures out the exact actions the car needs to take to drive safely. Finally, there is a module—the vehicle control and actuation module that takes those instructions and relays them to the car's physical parts on what to do, like pressing the gas pedal, applying the brakes, turning the steering wheel and using the turn signals. Like I said before, these modules are also data products.

Data products for decision support

Another very popular set of data products that can be widely adopted across the organization is the data products that provide the user with information to make informed decisions but do not take any decisions themselves. For example, customer engagement channel preference data products are designed to analyse and understand customer preferences for various engagement channels, such as online, mobile or physical visits. The insights generated are made available to decision-makers through interactive dashboards, making it easier to interact with the data and base decisions on it. While all the heavy lifting around cleaning, analysing and presenting the analysis is taken care of by this data product, the users still have to interpret the results and can choose to act on it or not.

The other three data products as I mentioned are more for the technical users and most often these are internal products. These data products can also become an integral component to building the two business focused data products.

Algorithms as data products

Algorithms are a set of mathematical rules and logics which take data as an input and generate valuable data or insights as an output. For example, a regression algorithm which is used to predict the behaviour of a dependent variable 'y', based on one or more independent variables 'x'. Repeatable algorithms used for multiple business use cases can be structured as a product. These products are anchored on algorithms that apply various analytical techniques, such as statistical analysis, machine learning or predictive modelling, etc.

These data products are typically integrated into the broader ecosystem through APIs, which allows the algorithm to receive data inputs, process them and deliver outputs or insights to other data products or users. Most of these data products act as an input for other more sophisticated data products, like the ones that automates action or provides decision support.

Algorithms are one of the key foundations for AI (along with data and computing power) and are often the ‘secret sauce’ as I say in Chapter 10, Proprietary Data. For example, this is how Google Image search works. When a user uploads an image, the product employs a sophisticated proprietary algorithm to analyse and understand the image. This involves extracting distinctive features, categorizing the image’s content and then comparing it to the vast database of stored images. The result is a selection of images that closely resemble the uploaded one, offering users an effective means to explore and discover visual content across the web.⁹

Derived and processed data as data products

The derived data products are built on the data derived from or based on raw data. The underlying data for these data sets are built by performing some basic level of analysis on raw data and only then are made available to the users through APIs. For example, a customer data hub that provides a comprehensive view of customers across businesses or sales data of various regions or products, etc.

Processed data is the most basic type of data product where the data is cleaned, transformed, aggregated and enriched to ensure it is made available in standardized usable format. For example, a customer demographics data set of an organization.

These data products are often used as inputs into or are used in conjunction with other data products.

Capturing that which is elusive

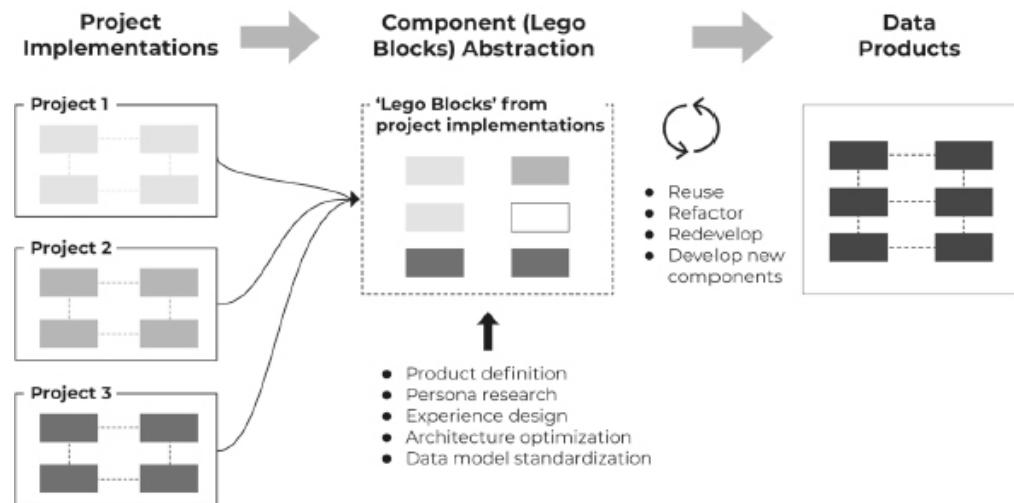
While we see the significant benefits of building data products for an organization, it is tricky to link data, insights and action in such a way that it achieves the desired impact. Since every organization has its own processes, structures and ecosystems, the one-size-fits-all approach does not work here either. So, building a data product is a complex and challenging task that requires careful planning, technical expertise and a deep understanding of both the data and the problem it is expected to solve.

In my experience, the most effective way of building and managing data products is to translate certain successful repeatable projects into data products. And here is the best approach to do so:

1. **Identifying repeatable use cases:** Data products are designed to achieve specific business goals and solve problems repeatable in nature by bringing together capabilities across the data stack. So, identifying these recurring scenarios in organizations is a necessary foundation for creating data products. This is a crucial first step because, if not done well, organizations may waste time and money developing data products that do not have a significant impact or that are not adopted by users. So, starting with the end in mind and establishing the target outcome is essential to ensuring that right use cases are identified for data products.
2. **Building a 'Lego block' inventory:** Once the repeatable need is identified, create a comprehensive inventory by cataloguing and documenting reusable solution components or modules from previous projects. This inventory can serve as a repository of tested components that have solved a problem or a use case. From this laundry list, teams can then easily identify and reuse existing solutions which can help in building the desired data products.
3. **'Lego block' prioritization:** This process involves careful evaluation of solution components on four factors —reuse, refactoring, redevelopment or creating new features to build the desired data product. Reusable components are identified and reused wherever appropriate, while others may require refactoring to optimize performance or meet new requirements. Some components may need redevelopment to align with updated technologies or architecture, and new features can be developed for the desired data products. This prioritization approach ensures efficient utilization of

existing assets while accommodating new requirements from the data products.

'Projects to Products' approach helps in converting specific projects into repeatable Data Products



4. Adopt a modularized, adaptive architecture:

Breaking down the solution into modular components enables easier adaptability, maintenance, updates and enhancements. So, the choice of frameworks and technologies for the data product should be made in such a way that they facilitate scalability and adaptability to enable seamless integration across different layers of the stack, across multiple processes and new features and functionalities as the product evolves.

5. Focus on user research and experience design:

By prioritizing user research and experience design, the focus shifts from project-specific solutions to outcomes or use cases, which requires understanding and addressing the needs of users across different projects. This process would enable building user-friendly data products that can enable better adoption and greater effectiveness with data.

6. Product manager-led multi-skilled 'pods' to own the product:

Often referred to as 'pods', a product manager-led team that has multi-skilled team members

including product management, data engineering, data science, as well as domain or business experts, enhances accountability and cross-functional collaboration. The product manager acts as the orchestrator, responsible for driving the product vision, managing priorities, coordinating efforts and ensuring the successful delivery of the data product.

7. **Feedback, learning and collaboration with end-users:**

Continuous feedback and collaboration with end-users are critical for the ongoing evolution of the data product. Regular user testing, surveys and feedback sessions provide valuable insights into user needs, pain points and opportunities for improvement. This feedback loop helps in identifying areas for enhancement, refining user experiences and prioritizing future feature development.

What does the future look like?

With the continuous advancement in technologies, growing availability of data and the evolving needs of a business, the evolution of data products looks both exciting and transformative. There are certain key aspects that would play a critical role in the evolution:

- 1. AI-powered data products:** In the future, I expect a convergence of technology, AI, data and domain expertise, leading to a unified ecosystem of data products. These AI-powered data products will utilize ML algorithms to extract valuable insights from large data sets, enabling accurate predictions and intelligent decision-making. For example, an AI-powered data product that leverages Fast Healthcare Interoperability Resources (FHIR) data sets can identify patterns, detect anomalies and provide personalized recommendations for diagnosis, treatment and disease management. Through

advanced analytics and ML algorithms, these products can enable healthcare providers to make evidence-based decisions, improve accuracy in diagnosis, enhance treatment planning, facilitate early intervention, reduce medical errors and ultimately lead to better patient care and improved health outcomes.

2. **Real-time and streaming analytics:** As the demand for real-time insights and actions continues to grow, so will the need for real-time data products that can process and analyse streaming data on the fly. Data products that leverage real-time and streaming analytics will enable organizations to respond to business situations promptly. A fitting example of this are the real-time and streaming analytics data products used to prevent money laundering in banks. These data products can continuously analyse large volumes of data from various sources, such as transaction records, customer profiles and external data feeds to identify suspicious activities in real-time. It enables organizations to take actions in real time such as flagging suspicious transactions for investigation, freezing accounts or initiating regulatory reporting.
3. **IoT generated data:** With the proliferation of IoT devices, source like sensors, wearables and connected devices will become a major source of incoming data. Data products will leverage more and more of this IoT-generated data to provide valuable insights. One such example is the preventive maintenance system of a coffee machine that collects data from various sensors that track aspects like temperature, pressure, usage patterns, etc. This data product enables real-time tracking of the machine to foresee potential failure or maintenance needs and tackle them in advance.
4. **Key to data democratization:** Data products hold the key to enabling self-serve capabilities to a wider number of employees across all levels in the organization,

through well-integrated pre-built assets that enable them to independently access and analyse data, eliminating the reliance on specialized technical skills. For instance, an AI-powered data product that leverages internal HR and finance records for employees, to provide a self-service tool that enables employees to access and analyse their own performance, compensation and career-growth metrics. This data product allows individuals to make data-driven decisions, such as identifying skill gaps, setting performance goals and seeking development opportunities.

5. **Tailor-made to industry or domain:** Data products will increasingly be tailored to specific industries and domains, specific to the unique business environment, regulatory requirements and challenges of each industry. Tailoring data products to specific industries allows for focused development and refinement of use cases that are most relevant and impactful for that particular industry. For example, a pharmaceutical company can leverage tailor-made data products to enhance patient medication adherence. These data products can analyse patient data collected through various channels, such as electronic health records and wearable devices, to generate valuable insights on medication adherence patterns and factors influencing non-adherence. This deep understanding of patient behaviour can help identify the root causes of non-adherence to design targeted interventions, develop personalized patient support programmes and much more.
6. **Data monetization:** Data products will become key for organizations to monetize data. Data products provide a way for organizations to package and deliver insights, analytics and actionable information. By transforming raw data into valuable products, organizations can create new revenue streams, unlock untapped market opportunities and enhance their competitive advantage.

A good example of this is Data as a Service (DaaS) for a banking firm specializing in mortgages. These data products can leverage their vast mortgage data to provide insights into customer behaviour, market trends and risk analyses which translate into improved customer targeting, personalized offerings and optimized risk assessment. Additionally, these data products allow the banking firm to streamline operations, enhance loan-approval processes and identify potential opportunities for growth.

Overall, data products hold tremendous potential to revolutionize industries, enhance decision-making processes, drive collaboration and innovate faster in the data-first world.

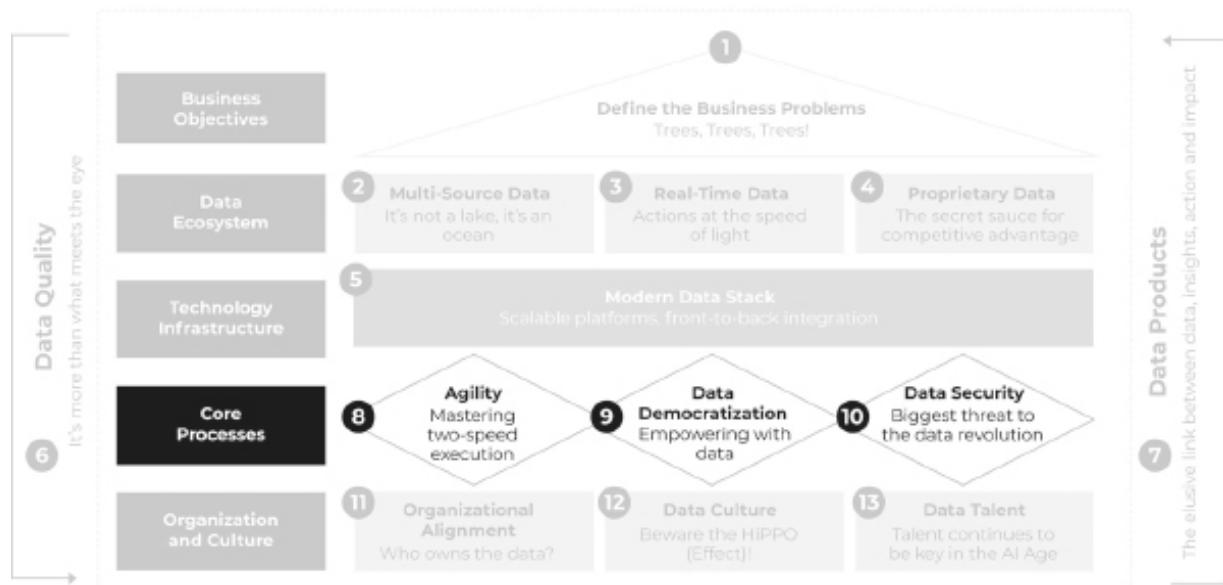
Key takeaways

- Data products are self-contained units built as a vertical slice, integrating various elements across the data stack to deliver business use cases in a repeatable manner and accelerate data to insights to actions to impact cycle.
- Data products have the potential to address many of the challenges in the Big Data world—ensure outcome focus from data, make the data-management cycle efficient, address ownership/collaboration issues and promote self-service.
- To effectively build and manage data products, organizations must start by identifying repeatable use cases and then prioritize the reusable ‘Lego block’ inventory of successful projects and adopt a modular, adaptable architecture. Additionally, they must establish product manager-led teams with multi-skilled members and maintain a feedback loop with end users for continuous improvement.

- Over time, boundaries between different product constructs, such as tech products, AI products, domain products and data products, will blur as any effective product will need to inherently incorporate all these elements.

LAYER 4

CORE PROCESSES



'Key to efficiency and effectiveness'

14

Agility

Mastering Two-Speed Execution

'In an era of predictable unpredictability, agility allows us to emerge in the here and now, while at the same time bridging our longer-term strategy with the present.'

—Roger Spitz,

The Definitive Guide to Thriving on Disruption: Volume II

Now that I have talked about what goes into building a modern data infrastructure, let me get into the core processes required for effective execution of data initiatives.

Here, let me introduce you to another paradox of the Big Data world, one of the many talked about in this book. One of the promises of Big Data is that it can enable quicker and effective decision-making, critical for organizations to create significant value, in terms of more revenue and reduced costs. However, the reality is that it takes a long time to establish the scaled-up, sophisticated solutions to enable such quick and effective decision-making. It also requires a significant amount of investment to go in before the solution can start generating any value.

While this is the traditional approach that most organizations take, there is a downside to taking too much time and investing too much money. The longer it takes to build these solutions, the chance of achieving the expected outcome, in a rapidly evolving and highly dynamic business environment, goes down exponentially. So, by the time the solution is built and ready, both the nature of the business problem and the outcome expectations may have evolved, and so the input in terms of data, tools and methods may need to be

revisited, rendering the solution obsolete. Only 29.2 per cent organizations, in a 2021 survey, reported achieving transformational business outcomes through their data and AI investments, while most reported struggling to make progress.¹

In this chapter, I will address this challenge and suggest an 'agile' approach to tackle it effectively.

The 'Big Bang' approach is bound to fail

The exponential growth in data generation from various digital sources has added to the complexity of dealing with such data. Consequently, the demand for more and more complex and sophisticated systems and applications that can deal with the vast and complex data is on the rise. As a result, they are investing in larger and more comprehensive data initiatives, with the ambition of harnessing the full potential of the data at their disposal. A 2021 C-suite survey, focused on the progress of Big Data and AI initiatives revealed that a staggering 91.9 per cent of companies are witnessing an accelerated pace of investment in such projects, with 62 per cent of firms disclosing investments exceeding \$50 million in data and AI initiatives.²

And as organizations continue to recognize the transformative potential of data, these initiatives are expected to keep getting bigger, becoming more 'Big Bang'. And therein lies the challenge. The bigger the size of the data initiative, more is the time and investment required to build it out. And the longer it takes, the higher are the chances of failure. By the way, this is not a one-off case, but a norm today. I have seen it happen time and again in my client situations.

This traditional Big Bang approach meant building a comprehensive, all-encompassing solution and large-scale deployment. This is often done by implementing all the necessary data, systems and processes in a sequential manner, often referred to as the waterfall method. Here the project is divided into distinct phases, with each phase being completed before the next one

begins. Carried out in a linear manner the progress on capability building is done in stages.

This Big Bang approach stems from a long-standing belief that 'bigger is better'. It is NOT! The approach, while emphasizing on comprehensive solutions, lacks tangible and immediate business outcomes in the short to medium term. But organizations no longer have the luxury to wait for all the stars to align. They must respond quickly, which requires building quick solutions. This is where the organizations have to change their game. Building solutions that deliver quick impact enables organizations to iterate fast, keeping pace with the dynamic business requirements and staying ahead of the curve.

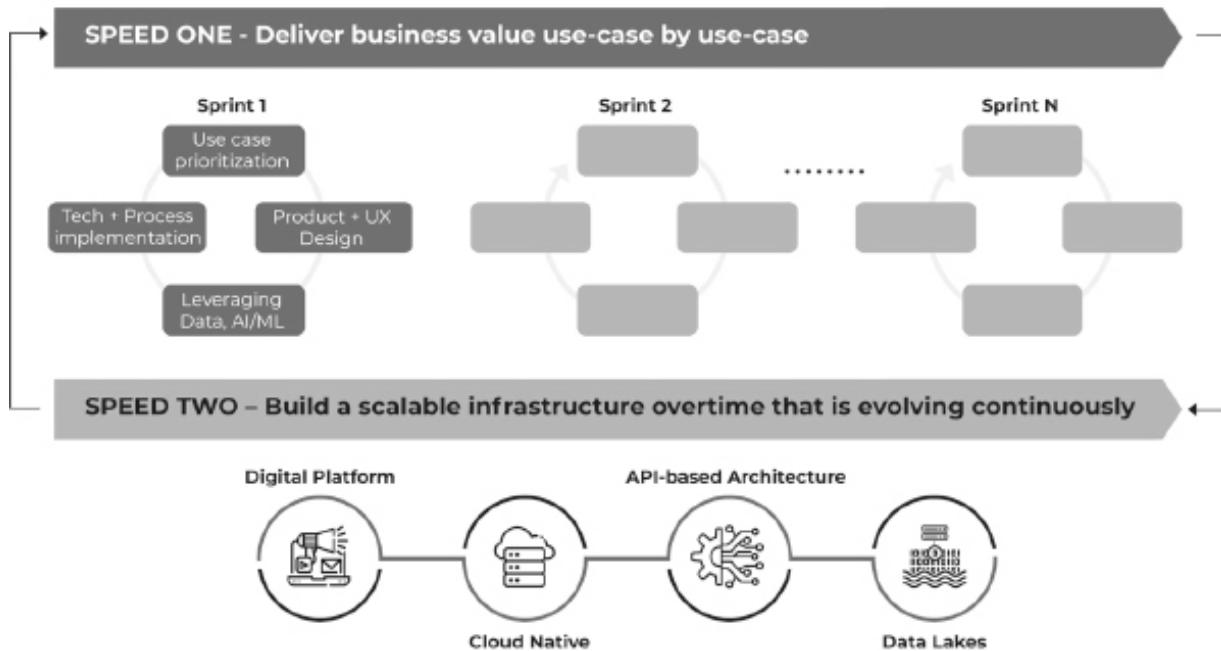
Having said that, focusing on building short-term solutions in a disconnected manner will not work either. Connecting these smaller pieces is equally critical to successfully build long-term capabilities.

Unfortunately, while 96 per cent of respondents in a survey underscore the criticality of agility for future success, only 26 per cent rated their company's current agility as high.³ Adopting an effective, agile approach would enable them to cater to the short-term business requirements while simultaneously building a scalable infrastructure for the long term. I call it the 'two-speed implementation' approach.

The two-speed approach

Although, Big Bang is no longer an effective option, focusing on narrow, short-term use cases would impact the scalability organizations require for the future. A more agile approach—which I call the 'two-speed approach'—focusing on delivering short-term business impact while ensuring that long-term capabilities are built in the process, would enable organizations to successfully build their data infrastructure. Let me now explain what the two speeds are.

Two-speed implementation is key to successful execution



Speed One: The high impact quick wins

To start with, organizations must identify high-impact use cases that solve an immediate or specific customer or business problem and deliver quick and direct impact. This can typically be done by quickly leveraging or connecting different layers of the data and tech stack. Since these are smaller, more manageable components, building a solution quickly is easier because data elements and pipelines that are required to address that particular problem are limited, as compared to bringing all the data together. It goes back to the concept of narrowing down the business problem, which I talk about in Chapter 7, Define the Business Problems. Identifying the most critical, actionable component(s) of the broader business problem, makes the whole exercise more manageable, bringing down the scope and complexity around the focus area and streamlining the data requirements, ensuring data quality is maintained throughout. Building a Speed One solution follows a similar approach to building a data product—it starts with the specific outcome in mind and is built by bringing all the layers of the data stack together, talked about in detail in Chapter 13, Data Products. But unlike a Data

Product, a Speed One solution is not necessarily repeatable and therefore not generalizable for broader use.

The minimum viable product (MVP)

An effective way of building Speed One is to develop an MVP that focuses on addressing a specific, high-impact business problem or opportunity. This allows for rapid prototyping and testing of ideas, enabling organizations to quickly assess the feasibility and value of a solution while minimizing upfront investment.

While there are numerous problems that an organization needs to address at any given point, it has limited resources available to tackle them all. Therefore, organizations must identify those high-impact use cases that are critical and can be translated into a minimum feasible solution to deliver value. Various existing components of the data and tech stack are leveraged to build this solution called the minimum viable product (MVP). While MVP focuses on a specific Speed One problem, they ideally should be developed keeping in mind the outcomes expected to be achieved with the Speed Two capabilities (more on this in the following section). For prioritization, organizations can evaluate the use cases on following four criteria to identify those that can be translated into MVPs:

1. **Problem relevance:** Evaluating the relevance of a problem involves considering its alignment with the organization's overall strategic goals and immediate business needs. It also includes assessing the potential financial benefits that solving the problem can bring, like cost-savings, revenue-generation or efficiency-improvements. Another important aspect to consider is the impact on multiple layers of the organizational stack.
2. **Complexity:** Evaluating the complexity of a problem involves assessing the time and cost required for implementation. This entails assessing the complexity around data involved, the implementation process and

the necessary changes required in the data infrastructure. Additionally implementing complex solutions would require realigning the processes, workflows and people involved, which is also a key consideration.

3. **Ability to scale up:** Organizations should assess whether the solution is a function-specific use case that addresses a specific business need or if it has the potential to be adopted across the entire enterprise. Striking the right balance between a solution that is neither too narrow nor too broad is crucial.
4. **Organizational maturity:** Organizations must assess whether their current technology infrastructure is capable of supporting the solution to the problem or if upgrades or integration efforts are necessary.

Speed Two: Building scalable, long-term capabilities

While organizations are building the Speed One capabilities, they need to build long-term capabilities too, necessary to scale up efficiently. Only focusing on narrow use cases would mean organizations will always be playing defence, unable to make a significant impact in the long run. They must therefore aim towards building long-term capabilities by bringing together the components that are built in Speed One, over time, through collaborative discovery, reuse of projects, cross-business-unit cataloguing and sharing of artefacts. Speed Two is like creating a connected network of various solutions which are built as part of Speed One, providing a more systematic and focused approach towards building that long-term capability.

There are certain key elements that typically comprise Speed Two. One is the **data architecture blueprint** that works as a guide in identifying and prioritizing Speed One initiatives, which can be brought together to build Speed Two. While this architecture acts as an effective guide towards building the long-term capability, the organization should not treat it as the ultimate one; rather it should

act more like the North Star. Because spending too much time designing the architecture can easily become an unending exercise. Rather, the entire process should leave room to learn from the Speed One initiatives and continue to iterate as required. Another important aspect to focus on are all the **data sources** required to build Speed Two. It is important to identify what data would go into building the Speed Two and what sources would the organization need to tap into to build it. They should also possess a solid understanding of the **data stack** necessary to support the required capability. I have already talked about the various layers of the stack in detail in Chapter 11, Modern Data Stack.

It's not easy to connect Speed One and Speed Two

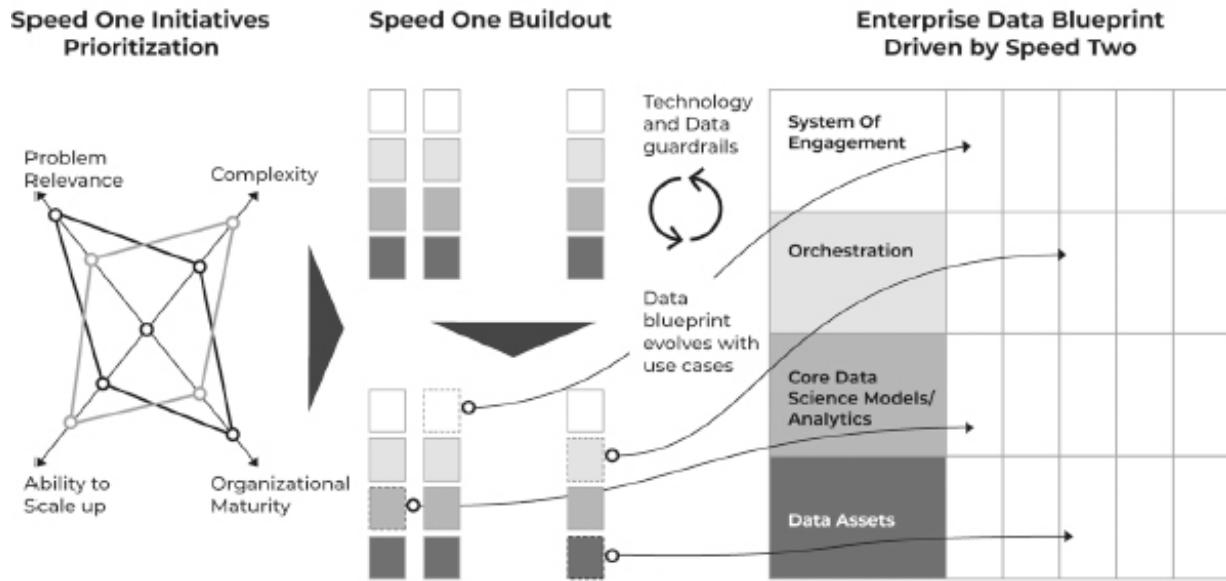
Ideally, Speed Two should be built by bringing together Speed One use cases. But it is not as simple as that. Because the nature of Speed One and Speed Two is so different, they are often of conflicting nature. Speed One focuses on quick wins, while Speed Two is about long-term scalability. Organizations often face difficulties reconciling these two approaches and end up running them as parallel efforts. Focusing on Speed Two leads to a delay in achieving business outcomes, while building Speed One(s) alone can mean neglecting the scalability aspect. Therefore, it takes a lot of experience and concentrated efforts to build Speed Two by leveraging Speed One use cases.

To tackle this effectively, organizations should start by establishing a guiding data architecture blueprint—a unique combination of data assets, data processes and data architecture—that aligns with their specific business needs. This architecture acts as a guide that evolves with changing business priorities and learnings from developing and executing Speed One initiatives. This blueprint serves as a foundation for building Speed Two from Speed One use cases.

The process involves evaluating MVPs for their reusability and potential to enhance the data architecture blueprint. Components that align with the blueprint can be incorporated; while those that

can't be, are discarded. Over time, this iterative process results in a scalable data infrastructure that supports the organization's evolving needs.

Capability building through two-speed approach



It is like curating a personal library at home. You start by collecting books that align with your interests and long-term knowledge goals. As you read, you assess whether the book is worth keeping based on its relevance. Some become favourites, while others are less so.

You then organize your library, by grouping and stacking books by topics or genres, creating comprehensive knowledge sections. Those that don't fit into any category are placed in a 'miscellaneous' section, which over time may evolve into a new genre as you add to or refine your collection.

The two-speed approach to building a customer 360-degree platform: An illustration

Let me take an example to bring the two-speed approach to life and make it easier to understand. And since I talked about the enormity of the exercise of building a 360-degree customer platform in

Chapter 7, Define the Business Problems, here is a practical (read two-speed) approach to achieving a 360-degree customer platform also known as 'single view of customer' or a 'customer 360'.

Typically, organizations attempt to tackle this by diving straight into Speed Two, where they try to consolidate all possible customer data to create the platform, only to find themselves overwhelmed.

The better approach is to start with a clear identification of the desired outcomes, such as attracting new customers, cross-selling to existing ones, or improving customer retention. With these objectives in mind, the organization should identify high-impact Speed One initiatives that would help meet those business objectives, while developing the necessary data architecture blueprint for Speed Two in parallel. My Speed One philosophy emphasizes prioritizing components that provide immediate business impact, fostering an outcome-centric mindset and optimizing the data stack to deliver value.

Let's take the example of a retail bank aiming to build a customer-intelligence platform. Here is how the bank should approach it:

Identify and build Speed One

As part of Speed One, they can start by addressing a key business objective like customer churn—the number of customers who have stopped engaging with the brand and stopped purchasing their products. Now, for a large company, the scope of churn itself would be very large and may need to be further broken down. Practically, the bank will have to start by addressing churn in one product category in one region for a particular customer group. To evaluate the customer churn, the customer population should first be split by type of retail banking product: savings account, term deposit, mortgage, credit card, etc. Then the performance drivers for each of the products should be disaggregated, to identify areas of highest concern. Let's say, based on the product split, the credit card customers emerge as a key concern, contributing to 40 per cent churn for the bank. The bank should then focus on credit cards as a product for building the Speed One capability for solving churn.

The performance KPIs for credit card churns can further be decomposed and analysed to understand the impact of the sub-components on the overall churn—customer service, customer onboarding, user experience, product quality. Upon analysis of the correlations with these sub-components, the most impactful driver for customer churn can be identified.

Let's say user experience is identified as the most impactful driver. A deeper dive will help narrow down on most-critical root causes. The next step is to define key actionables to handle the most-critical customer experience-related root causes of churn and prioritize based on impact vs complexity, to help improve customer experience. Now that the high-impact, low-complexity use cases have been identified, the data requirements can be highly streamlined to make the process easily manageable.

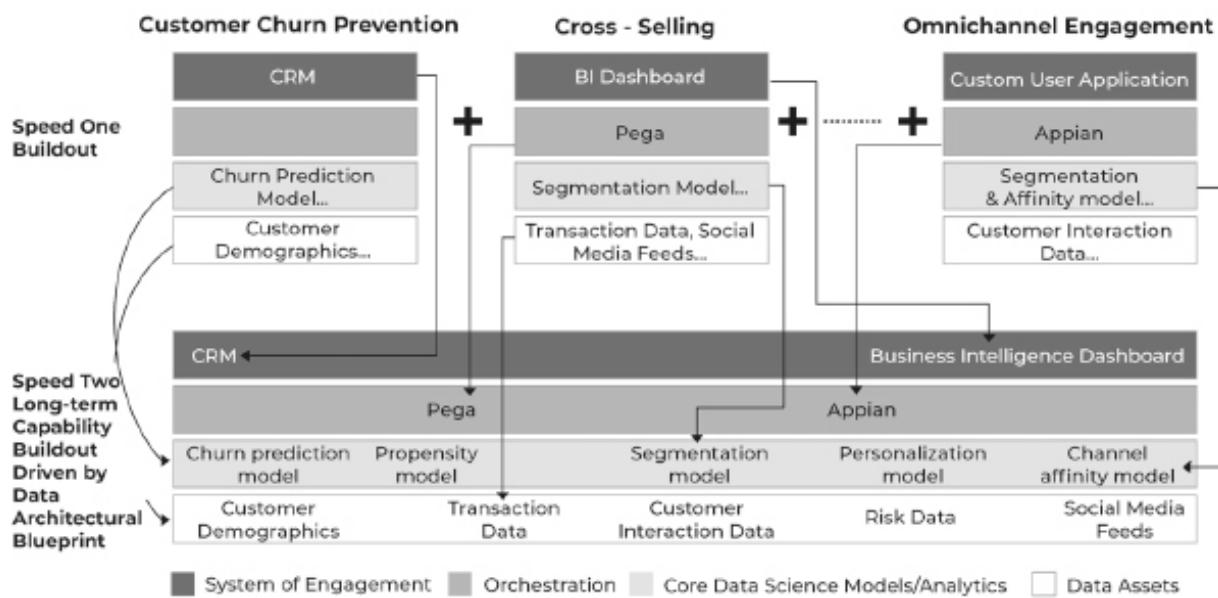
The various components across the data stack can be brought together accordingly, like specific data sources, the core data science models and the engagement layer that will help derive insights and take targeted actions to reduce customer churn. It also helps build proactive retention strategies, personalized customer interventions and the development of targeted retention initiatives to reduce churn and improve customer loyalty.

So, the solution to address churn is now built. But as I discussed in Chapter 13, Data Products, there is a possibility of dropout at each stage of the DIAI cycle, where the impact achieved is much less in comparison to the data we start with. Organizations can evaluate the solution at various levels. Against the objective of reducing churn, the organization can evaluate its performance across the DIAI—whether the data sources used are appropriate and complete, if the model is well-built, whether it is deployable or not, and once deployed, whether it is delivering the desired impact or not. Through this process, the organization can gather feedback at various stages of the DIAI framework, learn from it and make iterations to refine the Speed One as and when required to improve its effectiveness. These components can then be scaled across regions, products or customer segments to progressively build Speed One.

Another Speed One solution can be a cross-selling model—offering additional products or services to existing customers beyond their initial purchase to increase revenue and customer satisfaction. Again, specific data sources, when combined with appropriate core data science models, can help banks identify the most relevant cross-sell offers for individual customers, improve customer satisfaction, increase revenue and strengthen customer relationships.

Similarly, the bank continues to identify more such Speed One use cases that deliver high impact and can be quickly built to deliver value at speed and are aligned to the data architecture blueprint.

Case Example: Customer 360 platform



Build Speed Two data architecture blueprint in parallel

While the organization is working on building Speed One initiatives, they must also work towards building a guiding data architecture blueprint in parallel. For the customer 360 platform, the data architecture would typically look like this:

Foundational data layer: It involves identifying and defining the data sources that would become the data foundation for customer

360 platform. Relevant data sources need to be identified which typically includes customer demographic and socio-economic data, customer-transaction data (account transfers, loan payments, credit-card usage, Internet banking, ATM event stream), banking product usage data, response to marketing campaigns and social media feeds, etc.

Data science workbench layer: This layer provides the necessary tools and capabilities for developing and deploying advanced analytics and machine-learning models. This would commonly include models like customer segmentation engine, customer lifetime value (CLV) model, cross-product affinity, propensity scoring models, personalization model, churn prediction models, next best action recommendation model, etc. This layer should also have the ability to modify existing models and create new ones, as required.

Orchestration layer: This layer is crucial for delivering insights to internal and external user engagement systems. It focuses on automating the process of delivering insights to appropriate systems seamlessly through tools like Pega^{*} and Appian.[†]

System of engagement: The final and equally important layer focuses on data consumption and interaction with end users. It is that crucial link that empowers users to access and engage with the data and insights generated by the system. For a bank, this would typically include internal and external facing systems for the users such as core banking application, customer relationship management system, customer applications, marketing campaign orchestration system, decision support system—self-serve BI tools, etc.

Connecting Speed One for Speed Two

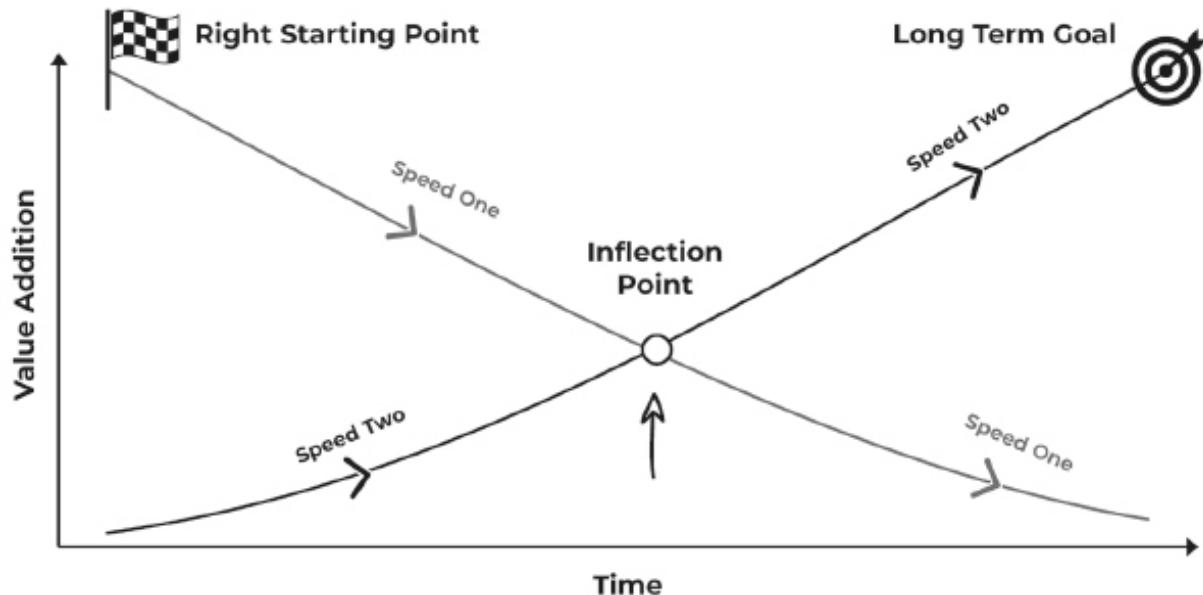
As various Speed One use cases are successfully built and deployed and are able to deliver the desired impact, like churn reduction for credit cards, churn reduction for term deposit, churn reduction for mortgage, etc., you start to develop a more comprehensive view of

the customers across different products by connecting the different data sets used for building the Speed One use cases.

As these Speed One use cases are built progressively, building the subsequent use cases would be relatively easier as some of the data sources, core analytics models would overlap with each other. For example, customer demographic data and customer segmentation model would be common for all customer-related data products. Also, the core data science model would get richer as more and more data sources are added. For example, customer segmentation model, personalization engine would get richer as we onboard data for different products.

These integrated solutions evolve as more and more use cases are added and help unearth some of the key insights on customers across various products. For example, opportunity to sell credit cards to customers who have a specific spending pattern on their savings bank account. And as you continue to bring different pieces together, the underlying customer data keeps expanding, becoming more comprehensive, enabling organizations to derive richer and more powerful insights, while the underlying data architecture continues to evolve.

Connecting Speed One initiatives to build Speed Two capabilities will help reach an inflection point where Speed Two capabilities start driving innovation



Over time, there comes an inflection point where data becomes a fuel for innovation. This is because as data continues to expand and grow, it attains a critical mass where it can be leveraged to uncover newer and more innovative insights. A good example here is how Google did it. In 1998, it started with a search engine that indexed the text on the internet and ranked the page based on relevance. This generated a massive amount of data. Over the years, Google progressively collected more data from multiple sources and leveraged it to create and refine multiple data products or solutions for its customers such as Google AdWords, Google Maps, Gmail. Through these solutions, it continued to acquire more and more data on its users globally. Over time, it reached a point where it was able to come up with newer and more innovative solutions by leveraging the underlying data which grew richer and more comprehensive.

One such example is Google Trends Platform, which started as an experimental prototype in Google Labs. It leveraged the massive data to identify trends in the searches on Google. To start with, Google used this as an internal tool to analyse the trends around

searches with the objective to optimize its ad strategy. As this product evolved significantly over the years, it has now become a go-to solution for a number of brands and marketers to optimize their SEO (search engine optimization) strategy.⁴

Key takeaways

- The traditional Big Bang approach to building data and technology infrastructure is problematic in the dynamic digital age. It can become such a mammoth task for organizations that by the time such a complex and sophisticated solution is built, it may become redundant or may not work at all.
- The two-speed approach enables organizations to focus on delivering short-term business impact while long-term capabilities are built in the process.
- Speed One is built by identifying high-impact use cases that solve an immediate or specific business problem through a minimum viable product (MVP) approach. Speed One is built by starting with a clear objective and then building backwards by bringing all the layers of the data stack together, to deliver on a specific objective.
- Speed Two is created by connecting Speed One use cases, to build a scalable long-term capability, keeping in mind the data architecture blueprint that acts as a guide to building the long-term capability.
- As Speed One capabilities are continually brought together in such a systematic manner, the underlying data expands to reach an inflection point where organizations are able to leverage it to generate newer and innovative insights.

Data Democratization

Empowering with Data

'We have an opportunity for everyone in the world to have access to all the world's information. This has never before been possible ... It's a tremendous equalizer. Information is power.'

—Eric Schmidt,

Co-founder of Schmidt Futures, former CEO and chairman, Google

For any organization to become truly data-driven, data must be effectively leveraged in decision-making at all levels in the organization. For this, data should be made available and easily accessible to all. It does not mean lifting all barriers and making all critical data available to each and every one at all times. Rather, it is about making the right data available to the right people at the right time and empowering them with the right set of tools and technologies to leverage it for decision-making. And that is the art of democratizing data.

However, data democratization often requires a radical shift in the traditional mindset, where data was hoarded by a few, considered as the source of power and hence source of influence/differentiation in the organization.

In this chapter, I will delve into the reasons why the real power of data lies in democratizing it, and suggest some key considerations to be kept in mind while doing so.

Historically, data was a source of power concentrated in the hands of a few

Traditionally, many organizations operated in silos and the data captured or generated by one function was owned tightly by a data owner. Most of these data owners were reluctant to share their data because it was considered a source of POWER! Since data was scarce and the ability to leverage such unique and scarce pieces of information was what set them apart from the others, the data owners were protective towards their data. They also feared losing control of their valuable data. Additionally, sharing of data was considered to be too risky because of privacy and security challenges. Concerns around misuse or misrepresentation of their data made data owners hold onto it tightly. This added to challenges around data sharing, which often involved a tedious process involving detailed business justifications and multiple approvals.

In addition to that, most business owners were, and still are, not technologically equipped to generate insights from the data. In a 2019 Deloitte study, 67 per cent of managers and executives reported that they were not comfortable accessing or using data from their analytics tools.¹ Hence, it was the job of a specialized team or a specific team member who had the technological capability to generate insights from the data. This created additional bottlenecks in the easy sharing of data between teams or functions. The business owners of different functions were required to raise a request with these specialists who may or may not have capacity to assist in time.

And lastly, since data was considered to be such a precious commodity, the data infrastructure was also designed with a narrow view keeping in mind the requirements of each team or function and building virtual walls to fortify it. This created a disconnected ecosystem of systems and high technological barriers to sharing of data between teams or functions, resulting in data silos.

Today, the power of data lies in democratizing it

The word ‘democracy’ comes from the combination of two Greek words. ‘Demos’ meaning people and ‘kratos’ meaning power. So, one can say that democracy truly means **power of the people**. For centuries, activists around the globe have rallied against the social and economic inequalities, fighting to take power away from the elite few to the many or all, demanding a society built on equal rights and opportunities.

Similarly, in the context of data, it was considered a source of power for ages, concentrated in the hands of few people. But the most prominent manifestation of the digital age, the internet, was built on the basic tenet of democratization of data—making information available to all. It is the great equalizer of the digital world, where information flows freely, transcending borders and boundaries. Today more than 5.3 billion users are connected to the internet worldwide, constantly generating and consuming data.² In the pre-digital era, when information was scarce and access to data was limited, organizations often had to rely heavily on consulting firms and external experts to provide them with even the basic insights. However, with the proliferation of data and the exponential growth in openly accessible data sources brought about by the internet, as discussed in Chapter 8, Multi-Source Data has empowered organizations with greater access to data, tools, knowledge and cost-effective solutions. One powerful example is the availability of open government data. Today, there is a vast amount of information on demographics, economics, healthcare, transportation and more, openly available on various government websites, which in the past was locked behind bureaucratic barriers, accessible only to a limited number of researchers or government officials. This data can now be easily leveraged by most organizations for various objectives like market research, risk assessment and compliance, strategic decisions like expansion, investment and more.

Similar is the case with internal data as well. The amount of data being generated by every function, every team is so huge that it is difficult for any one team to own or control it. And as mentioned in earlier chapters, today's fast paced business environment calls for high-velocity decision-making based on diverse internal and external data sources, which cannot be owned or controlled by any one function or team. Thus, hoarding data within specific teams or functions is impractical.

Additionally, in the digital age, achieving business objectives requires a collaborative approach to data and a fundamental shift towards decentralization. No single organization, function or team can achieve a business objective alone in today's complex and dynamic business environment. Cross-functional collaboration is imperative to generating deeper and comprehensive insights, and this implies more access and sharing of data across functions. Furthermore, to enable high-velocity decision-making, organizations must decentralize data access, empowering teams and individuals at various levels to leverage data for swift, informed decisions. Therefore, treating data as an enterprise-wide asset is crucial, enabling widespread collaboration and deeper insights to drive value.

Therefore, it is high time, organizations move away from the siloed mindset and democratize the use of data across all levels in the organization, without which it is impossible to build a truly data-driven organization.

Data democratization: Bringing the power of data to all

*Data democratization can be defined as 'the process of making data available to all employees **appropriately** with seamless, any time access while **empowering** them with the knowledge and tools to track and use it through self-service tools'.* In simpler terms, it is the ongoing process of providing everybody in an organization, irrespective of their technical know-how, access to required data, based on their persona or role, work with it comfortably, feel

confident talking about it, and as a result, make data-driven decisions every time.

Democratizing access to data lays the much-needed foundation to facilitate greater collaboration within and across teams and functions. Ease of data availability and access is key to promote and encourage sharing of data and building a strong sense of empowerment and trust among employees at all levels. As a result, it also helps organizations move away from the hierarchical culture of decision-making. Additionally, in the high-velocity decision-making environment that businesses are operating in today, easily and freely available data is also critical to enable real-time decision-making by bringing in speed and accuracy to the decision-making process.

What goes into democratizing data?

Amidst the various concepts of democracy, two key principles remain common: equality and individual autonomy. Similarly, data democratization is also predominantly designed based on these two principles: equality—levelling the playing field for both non-technical and technical decision-makers in the organization by making data easy to access and use—and individual autonomy—building awareness and providing self-service capabilities that enable every business user to take decisions and participate in problem-solving process.

There are three characteristics that define data democratization:

- 1. Data discoverability:** Every employee should know where to look for the data as and when it is required for decision-making. The required data should be easy to locate, should be in an understandable structure and lineage, and should be of optimum quality.
- 2. Data accessibility:** The second characteristic essential to data democratization is data access—tools to access the quality data. Once the data is located by the user, they require authorized ability to retrieve, modify, copy, or move data for the purpose of analysis and sharing.

3. **Data usability:** And lastly, users must be equipped with the self-serve capabilities to work with data and generate insights, based on the business requirements. Business users should be able to analyse the data in different ways, using multiple sources without requiring any external technical or non-technical help.

Having said that, there are certain challenges associated with each of these characteristics that makes data democratization a complicated process:

1. **Maintaining data discoverability:** Owing to the magnitude of data being generated and captured by organizations every day, it is not an easy task to keep track of what data is available and where it is available. Compounding this is the variety of data sources and the type of data being captured which adds to the complexity of dealing with it.
2. **Evolving data access:** As organizations are operating in a dynamic environment where the business problems keep evolving continuously, the data landscape is also changing and along with it the data requirements keep varying too. Organizations need to manage access, adding and removing access as and when required, all the while ensuring appropriate level of data security.
3. **Enhancing data usage:** The key to achieving success with data democratization is to ensure widespread adoption of self-service tools. This requires a focused effort towards ensuring continuous onboarding and training of every employee across various levels and identifying the right tools that work well with the current infrastructure and upgrading it wherever necessary.

To overcome these challenges, I recommend organizations must follow a systematic process for data democratization.

Data democratization is an iterative process

Data democratization is an ongoing process. And like any process, it needs to be continuously monitored and improved upon to drive maximum value for the organization. However, the road to data democratization is not as easy as it sounds. It requires a complete shift from the siloed mindset to a collaborative one. The steps to harnessing the full potential of data requires a systematic approach. In this section, I have laid down the process of enabling data democratization that can act as a blueprint for any organization that aims at achieving maximum synergies across various teams and functions.

The process of data democratization has three phases:

1. The planning phase: In this phase, organizations need to do an as-is to-be analysis to understand the best strategy towards data democratization. The key action in this phase is:

- **Understanding the data ecosystem:** Understanding the current data ecosystem will enable organizations to identify the data requirements based on the employee and their role and assess the bottlenecks in making the required data available for their persona. For example, the sales data of a company with customer details may be relevant for the after sales and service department, but the R&D function only requires the sales patterns essential to understand the trends and not the customer details. This will help create an access blueprint to ensure right access to right people and avoid unnecessary overload of data or accidental misuse of data across the organization.

2. The enablement phase: In this phase the bottlenecks around seamless sharing of data are removed and right

access is made available to users across the organization. The various actions in this phase are:

- **Unlock legacy/siloed data:** Once the blueprint for channelizing data has been developed, organizations must look at their legacy infrastructure to break down silos and integrate data across all systems. Legacy data systems must be either replaced or upgraded to seamlessly integrate with modern data management systems and cloud-based platforms to enable multiple functions to use the data and achieve the business objectives by taking decisions quickly and independently. The most effective way to break down the silos created by legacy infrastructure is to upgrade to cloud. This way the outdated, isolated systems are replaced with scalable and interconnected solutions essential for seamless data sharing and integration.
- **Enable persona-based access:** This is one of the critical steps to data democratization. Depending on the persona or the role of the user, organizations must design the level of access that each user is required to have, appropriate enough to assist them in their decision-making process. Once the access rights are determined, the right set of user-friendly tools should be made available to these users so that they can access the data irrespective of their technical or non-technical backgrounds. Therefore, these self-service tools should be simple by design, that is, designed to be as intuitive and user-friendly as possible. These tools should incorporate data and analytical features tailored to specific personas or user profiles, making the data manageable for the user.

3. **The scaling-up phase:** Once the bottlenecks are removed and right access is provided, the next phase is to facilitate widespread adoption of analytics and data

visualization tools by decision-makers at every level of the organization. This can be done by:

- **Enabling self-service:** Promote the use of self-serve tools—data analysis and visualization tools such as PowerBI,^{*} Google Data Studio,[†] data preparation tools such as Trifacta,[‡] Paxata,[§] or data exploration tools such as Qlik Sense,^{*} Looker,[†] etc., into the daily routine of employees. And to reinforce the importance of data, organizations must instil trust in enterprise data by using best-in-class data-management tools, integration solutions and data-quality software. Decision-makers at every level should be comfortable using self-serve tools in their day-to-day decision-making process, else efforts towards data democratization will fizzle out in the long run and organizations will have wasted huge amounts of money and time for nothing.
- **Build continuous awareness:** Most business users are not familiar with the analytical tools and dashboards and therefore are reluctant to use them in their daily routine. It is therefore important to build awareness and familiarity with these tools through effective onboarding and continuous training. All data users must understand and be comfortable with analytics and visualization tools to be able to fully leverage them.
- **Refine:** And finally, data democratization requires a continuous and iterative approach. So, organizations must continuously monitor their data ecosystem and identify gaps, if any. Evaluating the data sets and the tools for their effectiveness and level of adoption across all levels in the organization is an important barometer to gauge the progress made and make the course correction as and when required. Unless the efforts are being adopted organization-wide and are

measurably enhancing the process of decision-making at all levels, data democratization will not be a success. Through this process, organizations can continuously improve on their data democratization efforts and bring in the latest, more user-friendly tools that can enhance the end-user experience.

To make data democratization sustainable in the long run, organizations have to build the right culture, where employees across all levels in the organization consistently leverage data in their decision-making process—a topic that I will talk about in detail in Chapter 18, Data Culture.

Airbnb: The success story!

Like many successful start-ups, Airbnb has seen exponential growth as demand for its services grew at an unprecedented rate. Within a short span of seven years from its inception in 2007, Airbnb had surpassed major hotel chains as the preferred short-term rental accommodation provider. In the spring of 2014, it had about 10 million guests and over 550,000 listings worldwide.³ Today, Airbnb has more than 150 million worldwide users who have booked over a billion stays and have over 4 million Airbnb hosts listing their properties on its website.⁴

To support such phenomenal growth, Airbnb has been rapidly expanding its operations and hiring more and more people. In parallel, it saw explosive growth in the amount of data and the number of internal data resources like data tables, dashboards, reports, metrics definitions, etc. This created a key challenge in scaling the data-driven decision-making in the organization. As a result, there were numerous issues that employees were facing:

- Often teams had to rely on others to locate the right data resources.

- Absence of metadata, data lineage and context eroded trust in the data.
- Multiple data silos led to a disconnected data landscape.

This required a focused effort in democratizing access, building trust and enabling teams to use data for their decision-making process.

The 'Dataportal' emerged as a transformative tool designed to address these challenges by democratizing data access and fostering trust among Airbnb employees.⁵ It embraced all the three characteristics of data democratization that I have talked about:

Data discoverability: This tool enabled data resource search and discovery for employees in data exploration. It provided a unified search functionality across the entire data ecosystem, enhancing discoverability and surfacing metadata which provided detailed context for the data set. This tool created an integrated dataspace and provided a graph-based representation of the data ecosystem. This graph connected various resources, including data tables, dashboards, reports, users, teams and more.

Data accessibility: It provided a framework for best practices with data, providing guardrails where necessary and enabling easy access to data for specific teams based on their roles. It also streamlined popular resource sharing within teams. Additionally, it provides a consolidated view of all the data resources an employee has created, modified or consumed, and promotes transparency from both production and consumption standpoints.

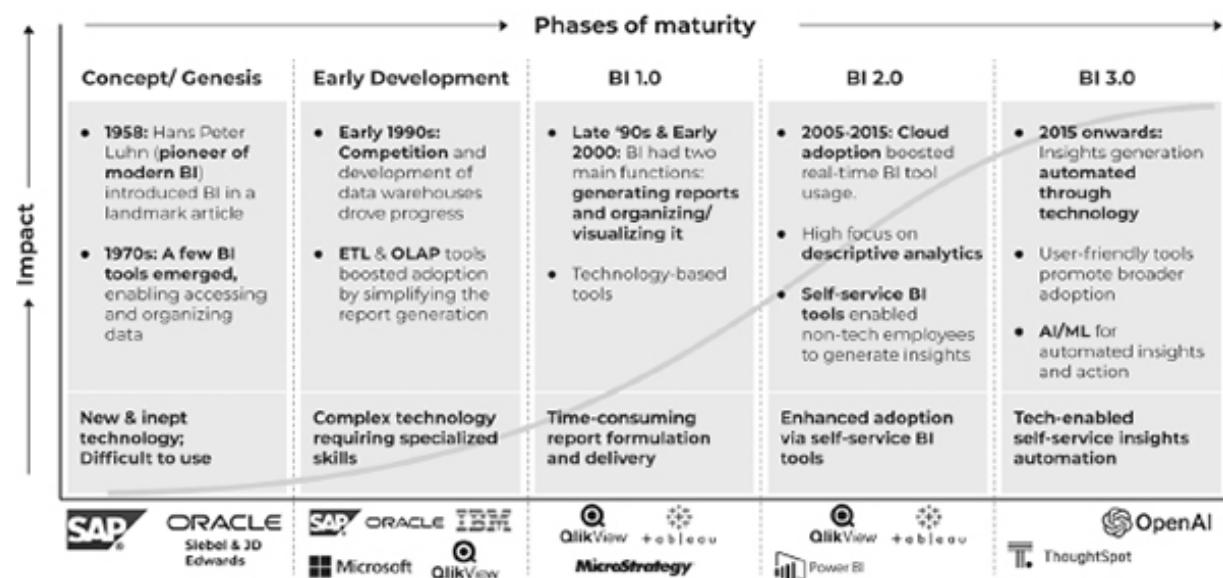
Data usability: Metric Explorer, a key component of Dataportal, enables out-of-the-box data exploration for teams across Airbnb. It enabled teams across Airbnb to easily slice and dice the data and make it easy to explore curated business metrics. Metric Explorer also supported powerful visualization capabilities which made it easier for teams to use data in their decision-making process.

'Self-service': A craftsperson is only as good as their tools

You can make all the data available, put in all the technologies required to analyse the most complex of data sets, but the real measure of success of your data democratization initiative is the extent to which your employees are comfortable in using that data. And because the majority of business users or decision-makers in an organization come from a non-technical background, it is extremely critical that the employees or business users at every level are comfortable with the data and are well equipped to analyse and visualize it with minimal assistance. This is where self-serve tools play a key role in maximizing the adoption of data-driven decision-making at all levels. Thanks to technology evolution, there are highly sophisticated BI tools available today to enable self-serve capabilities for non-technical users as well.

Initially, these tools were complex and very technical that required specialized skill sets. But today, BI tools are highly user-friendly, designed keeping the non-technical business user in mind. Let me take you through the interesting evolution journey of these BI tools and platforms that have made self-service possible.

Evolution of 'self-service' tools



The journey of self-service tools begins with the landmark article, written by Hans Peter Luhn, an IBM computer scientist, popularly known as the father of BI, describing the potential benefits of gathering business intelligence using technology. Over a decade later, there were a handful of extremely specialized individuals who were capable of translating data into useful information. And by the 1970s a few vendors emerged that made BI possible. But those tools were very nascent, clumsy and very difficult to use.

By early 1990s many more such vendors emerged increasing the level of competition and coupled with technological advancements like data warehouses, businesses started using in-house data analysis on a regular basis. The emergence of extract, transform and load (ETL) tools and online analytical processing (OLAP) further improved adoption. Still these tools were built on complex technologies that required specialized skill sets to work with.

Then came the era of high-tech BI tools. By the late 1990s to early 2000s, BI tools enabled basic functionalities of producing key analysis, building reports and organizing it in a presentable format. These tools were still not as user-friendly, and the process was time-consuming and tedious. By 2015, BI providers began providing simplified tools for decision-makers to become more self-sufficient. This was the era where more and more BI vendors started focusing on providing self-service as a key capability. Since the tools were easier to use and provided the functionality needed to generate insights, business users started widely adopting these tools and started working directly with data.⁶

Today, BI tools have become commonly used tools across every large and medium organizations irrespective of the industry they operate in. These tools work across multiple devices, like mobile, tablets and laptops, and apply analytical reasoning to data through interactive visual interfaces. And as organizations continue to generate and capture copious amounts of data, it is estimated that 95 per cent of new digital workloads—the amount of work that an organization needs to manage in a digital environment, will be deployed on cloud-native platforms.⁷ So, the on-premises BI tools

infrastructure solutions fall short to cater to the growing demand for flexibility and scalability required for higher network speeds and rapid shift towards virtual environments. Companies are therefore moving towards cloud-based BI and analytics platforms.

Furthermore, in the age of real-time analytics, live dashboards implemented across the organization help analyse, track and report the company's data in real time and are automatically updated to provide users instant access to critical data. In recent years, efforts are being made to move further ahead, moving from self-serve towards automation using augmented analytics—automating the process of generating insights. Analytics augmented with AI technologies such as ML and NLP, is a powerful way of combining human expertise with the power of artificial intelligence to provide insights that were previously unattainable. Using NLP can enable even an untrained user to query their data using voice or text commands to easily engage with the BI tools. For instance, Tableau offers a feature called 'Ask Data' that allows users to ask questions about their data in plain language, and it generates visualizations and insights based on the queries.⁸

Generative AI: The future of self-service

Tools like ChatGPT and Bard, are examples of how Gen AI has the potential to take self-service experience to the next level.

Gen AI can enhance the productivity and creativity of the user by generating customized and personalized content. With its conversational design, it can eliminate the technology barrier for non-business users. Since it is built on top of a large language model (LLM)—a very large neural network that uses deep learning techniques and massively large data sets and is trained to understand and generate human language, it can help uncover hidden trends and patterns that may escape human perception. For example, a recent study on the impact of Gen AI on productivity in the customer service sector, studied the impact of a chat assistant built using data from 5000 agents working for a Fortune 500

software firm that provides business process software. The tool, built on a recent model of GPT developed by OpenAI, monitors customer chats and provides agents with real-time suggestions for how to respond. It showed a worker productivity increase of 13.8 per cent, in terms of number of chats successfully resolved per hour.⁹ It also showed a disproportionate increase in the performance of less skilled, less experienced agents and enabling them to learn more quickly.

Another critical use case for Gen AI is the use of such tools in knowledge management. Enterprise search and knowledge management systems driven by conversational AI have the potential to democratize proprietary knowledge of the organization. Such use cases are already piloting in key sectors like healthcare, legal and financial services. Users can employ conversational language to search for information and receive more accurate and intuitive results and improve the overall search experience. It can facilitate personalized responses to inquiries by analysing vast amounts of information, identifying patterns and generating contextually relevant responses. It can help curate the knowledge base by analysing and categorizing vast volumes of information, to ensure that the most pertinent and current content is readily accessible to employees. It can process unstructured data, such as text documents, articles and even multimedia content, and extract valuable insights to enrich the knowledge base. Furthermore, it can adapt to user interactions, feedback and evolving data sources, continually refining its ability to generate accurate and useful information. And finally, it can be effective in continuously improving the accuracy and relevance of the information provided. As a result, the self-serve tools of today and those being developed are an effective way to reduce the employees' efforts and free up their time for more critical and higher value-add tasks.

Having said that, a significant hurdle to widespread adoption of Gen AI in their current states is the platforms' susceptibility to producing hallucinations—instances where the models generate inaccurate, faulty, misleading or even outright nonsensical

information. This tendency often arises when models are trained on limited data sets or contain false information. It can impact user trust and therefore needs to be mitigated with the use of high-quality data, increased transparency and stringent quality control measures. I have talked about how to do so in detail in Chapter 12, Data Quality.

Data governance is paramount

While I talk about making the right data available to the right people, at the right time, I cannot sign off on this topic without touching up on a foundational imperative to establishing robust data democracy in an organization—data governance (practices and processes that organizations use to manage, protect and optimize their data assets). Because, just like a democracy, we cannot have a thriving system without guardrails. Without some level of caution and restraints it would not take much time for it to turn into mobocracy—rule of the mob.

Robust data governance framework is a key enabler for successful adoption of data democratization. It is not just about extending data use and value across the organization and its ecosystem, but also about protecting it and meeting regulatory obligations. Without proper governance, data democratization can lead to chaos, data misuse and erode trust in the data. Data governance makes data accessible, usable and valuable, while ensuring safety and security. Here is how effective data governance benefits an organization:

Balancing access and control: Striking the delicate balance between giving broader access to data (data democratization) and maintaining control (security) is crucial. It ensures data is accessible to those who need it and shielded from unauthorized access or misuse.

Data collaboration: Data governance provides an effective framework for data management through robust standards for data

sharing, fostering a collaborative environment where data is trusted and data-driven decisions are made collectively.

Data-quality assurance: Data democratization is only effective when users can trust the data they access. Data governance also helps establish effective policies to ensure data quality throughout the data management value chain, enhancing the reliability of the data.

Data discovery and cataloguing: Data-governance practices such as metadata management and data cataloguing play a crucial role in making data discoverable and understandable. These practices are fundamental to data democratization, making it easier for users to find and comprehend the data they want to use.

Data-feedback loops: Data governance enables a feedback loop by providing a structured framework for data management and monitoring. Within this framework, users can engage with data, identify issues or discrepancies, and report their findings. These observations and feedback can be channelled back into the data governance process, leading to improvements in data quality, relevance and accessibility.

Data governance is becoming more dynamic and complex as organizations grapple with the challenges and opportunities presented by the evolving Big Data explosion, new technologies, regulations and work practices. On technology, some of the key drivers are cloud computing, AI becoming ubiquitous and increasing relevance of blockchain and decentralized data. On regulations—introduction of stringent data privacy regulations like GDPR (general data protection regulation) and CCPA (California Consumer Privacy Act), are forcing enterprises to upgrade their data management and compliance practices. And finally remote and hybrid working is also forcing a rethink on data management and control practices.

While we celebrate data democratization and other abundant opportunities of the data-first world, the focus on risk and control

issues is also increasing. We will explore a related topic, Data Security, in the upcoming chapter!

Key takeaways

- Data democratization is the process of making data available to all employees appropriately, with seamless, anytime access while empowering them with the knowledge and tools to track and use it for self-serve analytics.
- The key characteristics of democratizing data include making data easily discoverable, accessible with proper authorization and usable for analysis by all employees, promoting informed decision-making and reducing reliance on technical assistance.
- Data democratization is an ongoing process which requires a shift from the siloed to collaborative mindset, necessitating a systematic approach to harnessing the full potential of data.
- User-friendly business intelligence (BI) tools play a critical role to enable self-serve, empowering even the non-technical users to leverage data for decision-making. AI-augmented analytics further enhance accessibility and insights.
- Generative AI, exemplified by tools like ChatGPT and Bard, holds the potential to transform self-service experiences by providing customized content, uncovering hidden trends and democratizing knowledge management.
- Data governance is an important enabler for successful data democratization within an organization. It provides the necessary framework to ensure that while data becomes more accessible and valuable to users, it is also protected and compliant with regulations.

Data Security

Biggest Threat to the Data Revolution

'There are only two types of organizations: Those that have been hacked and those that don't know it yet!'

—John Chambers,
Former executive chairman, Cisco

In the previous chapter, I talked about the importance and enormous advantages of democratizing data, making data available for everyone in the organization to leverage for decision-making and driving actions. The flip side of doing so is that it also means opening up access to the organization's classified or critical data to more and more people through multiple touch points.

When you open all the doors and windows of a house for a gathering or party to allow easy access, there's a higher risk of unknown individuals entering unnoticed and potentially compromising the safety, security, and order of your home, belongings and the people inside. And that is why, in this Big Data world, where multiple stakeholders are creating and consuming the data across the organization's value chain, from across multiple touchpoints, the biggest threat that emerges is 'data security'. And as this ecosystem gets bigger, it becomes more complex, making it harder to safeguard the data. As a result, organizations are not only becoming more vulnerable to data leakage leading to security and privacy threats, but the scope of these threats is also getting bigger. So, in this chapter, I will talk about the growing importance of data security, and how to balance it while democratizing the use of data.

What is data security?

Data security encompasses a set of practices and measures implemented to safeguard data from unauthorized access, modification or disclosure. In simple terms, it is a way to make sure that the data is safe from leakage, misuse or alterations, either intentional or unintentional, at all times.

The proliferation of digital technologies and data-driven processes have led to an increased risk of data breaches and cyberattacks. Therefore, data protection is a critical aspect of modern business operations. In general, the process of safeguarding data begins with identifying 'what' data requires protection through data classification. Categorizing data based on its sensitivity and potential impact if compromised, ensures that the most critical data receives the highest level of protection.

Then comes the techniques and technologies to ensure data security. Techniques such as encryption, access controls and firewalls protect data from potential breaches or unauthorized intrusions. Encryption transforms sensitive information into an unreadable format unless decrypted with proper keys, ensuring confidentiality. Access controls ensure that only authorized individuals or roles have permission to access data, reducing the risk of unauthorized access.

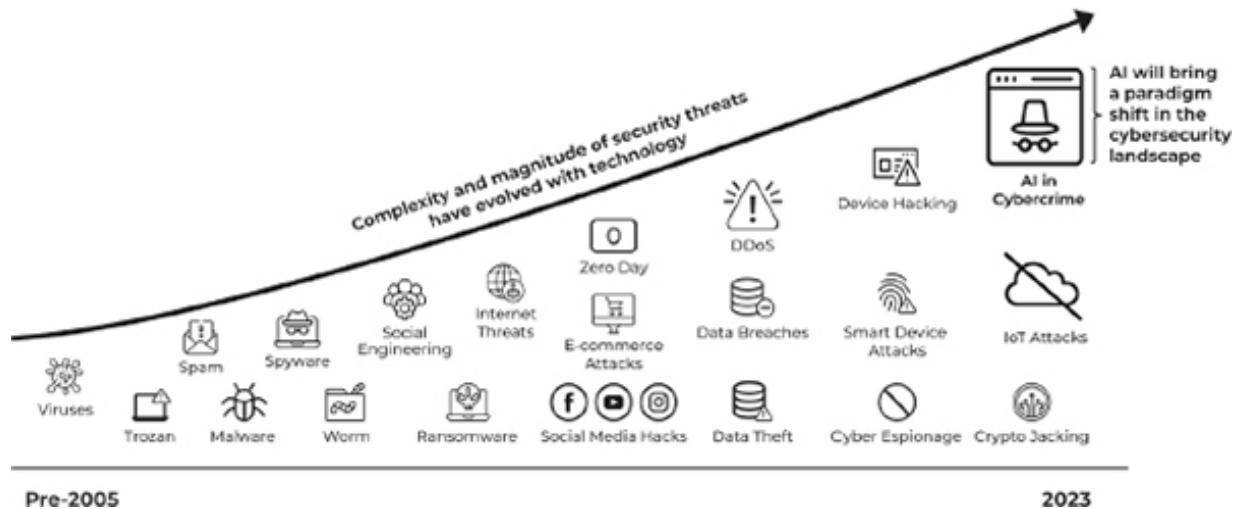
The primary objective of data security is to mitigate risks, maintain data integrity, preserve confidentiality and uphold the trustworthiness of data by preventing unauthorized access, disclosure, or misuse.

Data security: One of the biggest threats in the data-first world and AI age

In the past, security threats were significantly simpler and relatively contained owing to limited digital footprints, fewer interconnected systems and simpler attack vectors (the way they attack) like viruses, worms and basic malwares. But today we have come a long way from the relatively primitive forms of security threat like the first-ever cyberattack of the 1970s, often referred to as the 'Morris

'Worm', a computer program designed to exploit vulnerabilities in early Unix systems, that damaged around 6000 computers, and costed somewhere between \$1,00,000 and \$1 million to repair the effects of the worm.¹

Security landscape is evolving and becoming more complicated



Over the past few years, though, the proliferation of the internet, the rising adoption of digital technologies like cloud computing, mobile devices and IoT devices, have created a hyperconnected world. Particularly the significant shift from on-prem to cloud has created a more interconnected and open ecosystem. While this connectivity offers convenience, it has also expanded the attack surface, providing attackers with numerous entry points. Add to that, the exponential growth of data that has happened in the recent years, the magnitude of the potential for damage has grown tremendously. And as newer and more sophisticated technologies emerge, the nature of these threats is constantly evolving too, becoming highly complex and sophisticated! Complex attack vectors like multi-stage attacks or social engineering have emerged. Cyberattackers and hackers known as 'threat actors' who have evolved into organized cybercrime groups, hacktivists, etc., are using newer and more advanced technologies like AI and ML to launch targeted and adaptive attacks.

For instance, the Verkada attack, in March 2021, involved a group of hackers gaining unauthorized access to 150,000 cameras of Verkada, a company that develops cloud-based building security and operating systems.² This breach exposed sensitive video footage from various organizations, including schools, hospitals and businesses including Tesla, highlighting the security risks associated with IoT devices. Another example is the AI-assisted cyberattack of April 2018, where TaskRabbit,³ an online platform connecting freelancers and clients, experienced temporary suspension of the entire platform, affecting 3.75 million users who had their sensitive data compromised, including social security numbers and bank account details. Hackers, utilizing an AI-driven botnet, launched a devastating distributed denial-of-service (DDoS) attack on TaskRabbit's servers.

For organizations today, data is critical because it underpins virtually every aspect of their business, making it a highly valuable asset. Such complex cyberattacks can have a profound impact on organizations, manifesting in various detrimental ways. It can lead to resource misuse, which I closely witnessed in the recent past through what is called a cryptojacking incident. The hackers gained unauthorized access to the computing resources of an organization to mine cryptocurrencies like Bitcoin. While it did not necessarily lead to any sensitive information theft, the incident became an urgent concern for top executives, necessitating a cross-functional team—including not just security experts but delivery and business leaders—to spend many sleepless nights together on identifying and containing the breach and then remediating it. Globally, 332 million cryptojacking attacks were recorded in the first half of 2023, surging by about 399 per cent over the past year.⁴

On the other hand, cyber breaches can also lead to huge financial losses, often associated with ransomware attacks. Imagine if the above attack was made with the intent to sabotage all the systems where crypto mining was done. And the systems were held at ransom, encrypted and locked and made completely inoperable. A real-life instance from another organization illustrates this point

vividly. In this case, cyber criminals successfully infiltrated the organization's systems and gained control of its entire client database, encrypted it, exfiltrated the data, locking all systems, making them inoperable. They subsequently demanded a hefty ransom under the threat of releasing this sensitive information to the public. Often, organizations that find themselves in such situations have to hire specialists, like cybercrime experts, incident response team, digital forensics investigators, etc., to investigate and deal with such situations and end up paying tens of millions of dollars as ransom in the end. Not only that, it can also lead to severe damage to the brand image and reputation of the organization.

Security breaches can be internal, known as insider threats—caused by internal stakeholders like employees, contractors or business partners—or external—caused by malicious elements like hackers, etc. They can also be either intentional—owing to malicious intent—or unintentional—resulting from mistakes or oversight; or threat actors exploit human aspects of urgency, fear, greed and curiosity to bait them. But either way, the nature and magnitude of these breaches will continue to evolve as organizations deploy newer technologies and open their data up for more people in and out of the organization. It is estimated that the global cybercrime costs would increase by 15 per cent annually, reaching an astounding \$10.5 trillion per year by 2025.⁵ This cybercrime cost includes data damage and destruction; theft of money, intellectual property, personal and financial data, loss of productivity, fraud, embezzlement. It also includes post-attack business disruption, forensic investigation expenses, recovery and removal of hacked data and systems and damage to overall reputation.

Clearly cybersecurity is a CEO and board-level issue. It is astonishing how these issues rapidly escalate, often necessitating board-level involvement. And justifiably so, because, with such incidents, the value at stake is so high, that containing it becomes a matter of organizational survival, quite literally, especially when customer data is in jeopardy.

Historically, the major global financial crises have been driven predominantly by financial crime like in the 1995 Barings bank collapse or due to complex financial instruments which led to the 2008 financial crisis. While it might sound dystopian, I believe it's not far-fetched to anticipate that the next financial crisis could arise from a massive data breach. Unfortunately, the warning signs are already evident. For example, the infamous WannaCry Ransomware Attack in 2017, afflicting over 2,00,000 computers in over 150 countries, bringing the UK's National Health Service (NHS) to a complete standstill for days, and affecting many other companies including Spain-based Telefonica, America's FedEx, German railway company Deutsche Bahn and LATAM Airlines.⁶ Or the NotPetya cyberattack the same year that shook the world, quickly spread to more than sixty countries, infecting organizations in several sectors, including finance, transportation, energy, commercial facilities and healthcare, resulting in collective damages pegged at more than \$10 billion.⁷

Therefore, a critical concern for organizations today is 'how to keep their data safe and intact?'

The 'trust no one' policy: Zero trust

Traditionally, organizations have relied on the 'castle and moat'—a network security model where organizations establish a well-defined boundary around their network infrastructure, similar to a castle surrounded by a moat. It is a security model in which no one outside the organization network can access the data inside but anybody inside the network is inherently trusted. The castle and moat approach relied on the assumption that as long as the perimeter defences were strong, the organization's assets would remain secure.

Unfortunately, this approach to security is no longer effective for organizations.

The rapid adoption of cloud services, mobile devices, remote work and interconnected systems has blurred the lines of a traditional network perimeter. Organizations now operate in a highly dynamic

environment where data is decentralized across multiple platforms, devices and networks, making it challenging to establish a clear boundary for protection. Therefore, the focus has shifted towards implementing a more holistic and adaptive security approach that includes robust authentication mechanisms, encryption, continuous monitoring and threat intelligence to detect and respond to potential threats both within and outside the organization's perimeter—the core philosophy that assumes that the security of the organization's complex network is always at risk to external and internal threats.

So today, most organizations are moving towards zero-trust framework. This framework challenges the traditional idea of trusting everyone within the organization's network. In a zero-trust network, it is assumed that potential security threats can emerge from both inside and outside the network, eliminating automatic trust for users and machines. It follows the general principle of 'block all and allow only the necessary'. It also involves verifying every access request before granting it, ensuring a cautious approach to security.

In the context of the example of a house that I introduced in the beginning of the chapter, a zero-trust framework can be compared to a comprehensive home security system that ensures protection and verification at every point of entry or interaction within the house, regardless of whether it's from the inside or outside. Unlike the traditional house security setup, which gives people access to all rooms and resources once inside, a zero-trust framework takes a different approach. It is like having multiple layers of security measures inside the house as well, such as keyless entry, biometric authentication and surveillance cameras at every entry point. Each room within the house has its own access controls and permissions based on the principle of least privilege, meaning individuals only have access to the specific areas or resources they need. For example, a cook does not need access to the bedrooms in the house.

The zero-trust framework is an IT security model emphasizing strict identity verification for all individuals, devices or applications seeking network access, regardless of their location. It ensures authenticated users and devices receive customized access to

resources based on the user profiles, continuously verifying authorization and authentication throughout the network, unlike traditional models that authenticate users only at the network perimeter.

Organizations are increasingly acknowledging the significance of zero trust in ensuring robust security, which has resulted in a continued adoption trend. A study conducted in 2022 revealed that over half of the surveyed organizations (55 per cent) have already initiated zero-trust efforts, with an overwhelming majority (97 per cent) planning to implement such measures within the next twelve to eighteen months.⁸

However, when organizations operate with the philosophy of trust nothing and no one, it brings an important paradox to the forefront.

The security versus democratization paradox

The data security versus data democratization paradox refers to the inherent tension between the need to ensure data security while promoting seamless data access and sharing across an organization. On the one hand, data-security measures are necessary to protect sensitive information, prevent unauthorized access and mitigate the risks of data breaches. On the other hand, data democratization aims to provide broader access to data to individuals or teams, which requires breaking down data silos to promote seamless data sharing and enable self-service access to data.

Balance between Data Security and Data Democratization

Data Security

Protects the data from unauthorized access, modifications and prevent against data breaches



Data Democratization

Make data accessible to the right people at the right time and empower them with self-serve tools for decision-making



Balancing data security and democratization requires **enabling access while maintaining integrity and ensuring protection**

So how do you build walls and monitor every movement in the house without making it difficult and cumbersome for the members of the house to easily move around and go about doing the things they need to do?

Addressing the apparent paradox between data security and data democratization sounds like a difficult problem to decode. While it is not impossible to achieve, it certainly requires a balancing act—making data easy to share and use, while ensuring its integrity and security at all times.

Context is key!

Throughout this book, you may have observed that I have consistently highlighted the importance of keeping the context in mind while solving any business problem. Whether it is narrowing down the business problem or approaching data quality in the Big Data world or identifying the right use cases for real-time data—I have emphasized the importance of leading with the context. This approach is also key to simplifying the zero-trust implementation, ensuring a balance between security and democratization.

Taking a blanket approach of treating every user, device, application or network connection with the same lens will not provide an effective solution. Without the right context around aspects like the criticality of the data, its intended use, the user experience, among other factors, will result in sub-optimal security solutions which would become an impediment to democratizing data. By keeping the context in mind, organizations can streamline and tailor their security measures based on specific situations and risk factors. In simple terms, the context can be determined based on three critical factors: The data, the user and the activity.

The three key considerations to build a zero-trust framework

I believe that the most effective way to go about building the zero-trust framework, with a balance between democratization and security, is by focusing on the following three key considerations:

How critical is the data?

Of course, when the hero of the story is data—the element that we are looking to secure—it becomes critical that we start by evaluating the data itself. Simply put, not every data that flows in and out of the organization requires the same level of security control.

Data can vary in terms of its criticality and sensitivity. While it's important to ensure security for all types of data, certain data might be less critical to secure than others. For example, data that is publicly available on the organization's website such as general company information, press releases or product descriptions, etc. While this data should still be protected against unauthorized modification or impairment, it does not require security measures against theft.

On the contrary, personal identifiable information (PII), any data that can be used to identify an individual or distinguish one individual from another, like full name, date of birth, social security number, address, phone number, etc., is sensitive information and

requires special protection against exposure or misuse. While it is essential to collect such information for legitimate purposes, such as providing services or conducting business transactions, it is essential to handle PII responsibly and ensure that appropriate security measures are in place to prevent unauthorized access or breaches.

Through 'data classification', organizations must first categorize and segment their data based on factors such as criticality, sensitivity and usage. Based on this segmentation, it's easier to identify the more sensitive and critical data and choose appropriate levels of security accordingly. This approach also enables a more nuanced and efficient allocation of resources as the organization can prioritize their efforts as per the criticality of the data type.

Let me provide a very simple, practical example, which is true for any company, including Incedo. At Incedo, we handle different types of data on a daily basis, including employee data, company data and client data to name a few. The same level of security is not required for all the data in the organization. For example, banking credentials and transaction details of the firm are highly sensitive data and therefore require higher levels of security and greater access control. On the other hand, internal data which is low on criticality such as internal policies or employee handbook which has information that every employee has access to are not highly sensitive in nature. Here a lower level of security would suffice.

Who is using the data?

Customization based on the user profile or persona is another critical lever around which the zero-trust framework should be designed. It involves customizing access rights based on user roles and responsibilities. This would allow for more granular and tailored security policies to be designed that are aligned with the business's specific needs.

Instead of relying on static, one-size-fits-all access permissions, context-based access control ensures that users are granted the appropriate level of access based on their roles and responsibilities. For example, a user with administrative privileges accessing sensitive

financial data from their office workstation may have more restricted access compared to the same user accessing non-sensitive data from their personal device and a trusted network.

Identity and access management (IAM) solutions like role-based access control (RBAC) and context-based access control (CBAC) can be used to enable this. It is a cybersecurity discipline focused on managing user identities and access permissions on a computer network. The primary purpose of IAM is to facilitate easy and secure access to required data, based on individual persona or role in the organization. There are two parts to IAM: identity and access.

Identity management checks a login attempt against a constantly updated identity management database, which is an ongoing record of everyone who should have access. Access management is the second half, which once the identity has been confirmed, keeps track of which resources the person or thing has permission to access. Access to resources and data is subject to different levels of permission, which are determined based on factors such as job title, length of employment, security clearance and project involvement.

Moreover, the use of biometrics in authentication can enhance security while simplifying the authentication process. Biometric authentication, such as fingerprint or facial recognition, provides a seamless and user-friendly way to verify identity without the need for complex passwords or additional authentication factors.

For example, a retail bank employs IAM to safeguard customer data. IAM employs role-based permissions; customer relationship managers access data pertinent to their roles, like account and transaction details, and demographic information of the customers mapped to them. Whereas the back-office team requires authenticated access to customer account and transaction data. Any other access request would need an additional layer of security. In addition to the desired level of security, IAM solutions are also effective in ensuring high-quality user experience through simplified access procedures and password-less access.

What's the pattern of activity?

Another important consideration to base the zero-trust framework is the activity. Behaviour-based security systems, utilizing advanced techniques like analytics, ML and AI, detect abnormal or malicious behaviour by analysing everyday activities, such as logins, file access and network traffic. These systems collect and scrutinize a wide range of system and user actions, identifying suspicious patterns that may signify a security threat. For instance, they can raise alerts for sudden surges in failed login attempts or unusual network traffic. When an anomaly is detected, it triggers alerts for security analysts who then investigate to determine if it poses a threat and take appropriate action.

User and entity behaviour analytics (UBA/UEBA)

UBA or UEBA is a comprehensive defence mechanism against emerging cybersecurity threats. It is an effective tool which primarily focuses on meticulously scrutinizing the behaviour of both human users and non-human entities, such as devices (for example, routers and servers) and applications, with the goal of uncovering any irregularities or suspicious activities. Through continuous monitoring and pattern analysis, UEBA systems swiftly identify unusual behaviours and deviations from the norm, issuing alerts at the first sign of potential compromise. This comprehensive, real-time approach ensures the protection of sensitive data, operational integrity and resilience against a diverse range of threats. Most common example is when a user account is accessed from an unusual device, browser or location. In such cases, a security alert is triggered automatically which requires an additional layer of authentication.

Behaviour analytics can be categorized as external and internal, that enable an organization to proactively monitor and protect its entire digital infrastructure against various types of threats:

1. **Insider threat behaviour analytics (ITBA):** ITBA represents a critical subset within the broader domain of UBA. ITBA specifically focuses on identifying potentially malicious insiders—individuals who, by virtue of their roles or positions within an organization, possess trusted access to sensitive data and systems. By establishing baselines of normal behaviour, ITBA can detect deviations or anomalies that may signify insider threats. For example, it might flag an employee who suddenly accesses sensitive data unrelated to their job function or exhibits irregular working hours that deviate from their usual patterns.
2. **External behaviour analytic (EBA):** Sometimes referred to as external threat behaviour analytics (ETBA), EBA focuses on monitoring and analysing the behaviour of entities outside an organization's network or perimeter. EBA predominantly involves tracking the behaviour of external entities like websites, IP addresses or domain names that interact with an organization's systems. This approach helps organizations detect and respond to threats that originate from external sources, such as cyberattacks and malicious activities targeting the organization's online presence or digital assets. For example, detecting a DDoS (distributed denial-of-service) attack, by identifying a sudden and massive increase in incoming traffic from multiple external sources, allowing immediate intervention and remediation.

A behaviour-based security approach provides several advantages, notably **proactiveness** in detecting and addressing potential threats before the harm is done. It offers **comprehensiveness** through the process of monitoring a diverse array of system and user activities, providing a holistic view of security risks. Moreover, it exhibits **effectiveness** in identifying a wide spectrum of security threats, spanning insider threats, external risks, malware incursions

and data breaches, bolstering an organization's overall cybersecurity posture.

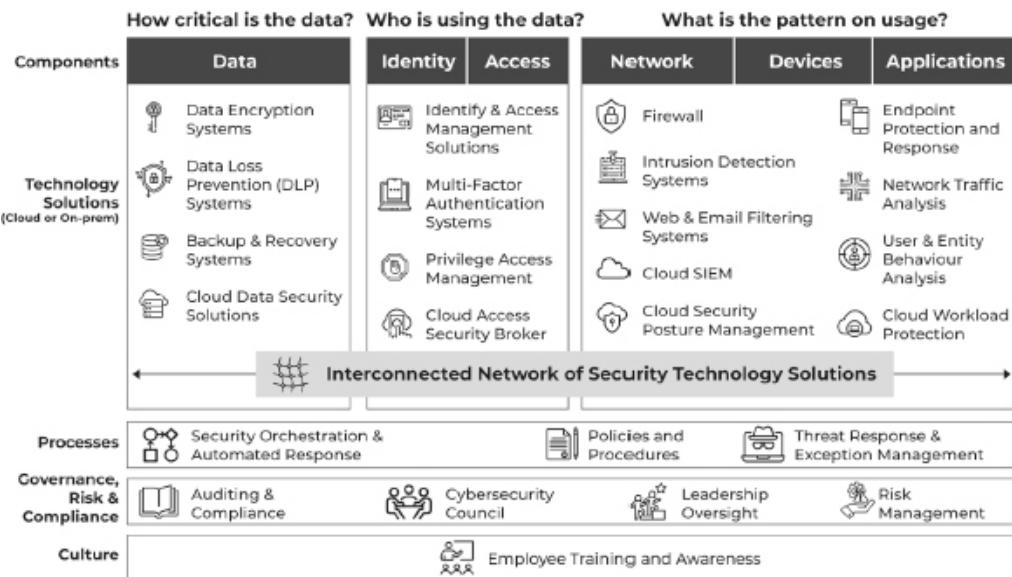
Key levers to execute the zero-trust framework

Now that I have talked about the three key considerations that go into establishing a robust zero-trust framework, let me delve into the various 'levers' that are critical for effective execution. While this topic in itself is so vast that it requires me to write a separate book, and maybe I will, but here I'll provide a concise overview and present a practical framework for implementing a context-driven zero-trust security framework in your organization.

There are four set of key levers that come into play:

1. The most critical lever is the availability of diverse **technology solutions** that are now available to choose from. Given the variability and complexity of the ever-expanding Big Data world, there is no one-size-fits-all technology solution to ensure comprehensive security. Depending on the specific context and business requirements, organizations may require to choose a specific product or solution, across the various layers of the data stack. However, the building blocks of technology solutions remain the same, and typically include firewalls (typically the first line of defence), data leakage prevention (DLP), endpoint protection, data encryption systems, identity and access management solutions, security information and event management (SIEM), user behaviour analysis systems, and more.

Building a context driven zero-trust infrastructure



With such wide and varied technology solutions available, the key is to interconnect all these systems. A potential solution can be to create an adaptive interconnected network of cybersecurity solutions, like a cybersecurity mesh, thereby enhancing the organization's ability to detect and respond to threats through the sharing of data and intelligence across its cybersecurity technologies.

2. Next comes well-defined and automated **cybersecurity processes**. Anchored in threat response and exception management mechanisms, policies and procedures, and security orchestration and automated response (SOAR), these processes ensure effective threat management. These processes help organizations to respond to security incidents, proactive threat hunting and improve their security landscape by providing a more comprehensive view of their security infrastructure quickly and effectively.
3. The next essential lever is a proactive approach to security with effective **governance, risk and compliance (GRC)** mechanisms which encompasses

various essential elements, including auditing, compliance with regulation and standards, risk management, establishment of a cybersecurity council and leadership oversight. Auditing and compliance mechanisms are crucial for ensuring that the organization adheres to industry standards and regulations. Regular audits and assessments help identify vulnerabilities and validate the effectiveness of security controls.

Concurrently, a dedicated cybersecurity council, comprising cross-functional stakeholders, provides valuable insights, input and oversight to steer the organization's cybersecurity strategy. Lastly, leadership oversight is paramount, as it ensures that senior executives are actively engaged and supportive of cybersecurity initiatives and it becomes a key agenda for CEO and board reviews.

4. Finally, building a **cybersecurity culture** within an organization, anchored in employee training and awareness, is fundamental to fortifying the organization's defences against cyber threats, in the long run. While I have dedicated an entire chapter on building a data-driven culture, Chapter 18, here I am highlighting a specific aspect of culture from a cybersecurity standpoint. It begins with comprehensive and ongoing cybersecurity training programmes that educate employees about the latest threats, safe online practices and the importance of vigilance. Regular awareness campaigns, gamification of phishing, simulated campaigns to develop social engineering awareness reinforce these principles, keeping cybersecurity top of mind for the entire organization is the key.

As is evident, establishing a zero-trust framework is not a one-time task but an ongoing and dynamic process. This approach acknowledges that the security landscape is constantly evolving, with new threats and vulnerabilities emerging regularly. Therefore,

organizations must continuously adapt and refine their zero-trust strategies to stay ahead of potential security risks.

AI is fundamentally transforming data security

AI can play a crucial role for data security due to its ability to analyse vast amounts of data, detect complex patterns and adapt to evolving threats in real time. With the exponential growth of data and the increasing sophistication of cyberattacks, traditional security measures alone are often insufficient. AI-powered systems can continuously monitor network traffic, user behaviours and system logs to identify anomalies indicative of potential security breaches, automating security tasks and improving incident response times significantly. Across the data security value chain, AI can **protect, detect, investigate and respond** to security threats with greater speed and precision:

Protect: Machine learning algorithms are used for **adaptive authentication**, to assess the risk associated with user activities based on behaviour and context. Adaptive authentication can detect suspicious activities and prompt for additional authentication or block access if required. For example, banks use intelligent adaptive authentication which asks for additional factors to prove identity like a security question or additional one-time password if the account is accessed from an unknown device. AI can also be used to create strong **encryption algorithms** that enhance data protection. These algorithms can transform sensitive data into unreadable format to protect it from unauthorized access. For instance, AI can generate complex encryption keys that are resistant to brute-force attacks—a hacking method that uses trial and error to crack passwords, login credentials and encryption keys, ensuring the confidentiality of data during transmission or storage. Additionally, AI can also help automatically identify and classify sensitive data, to define security measures appropriately.

Detect: AI can prove to be a powerful and proactive tool to detect potential security threats. AI algorithms can analyse **user behaviour patterns** and **identify anomalies** like unusual login times, access requests or unusual data access patterns. AI solutions with real-time monitoring capabilities can identify and detect intrusions as they occur, by analysing network traffic, system logs and other relevant data sources. For example, for a bank, AI algorithms can analyse large volumes of data, such as financial transactions or user activities, and compare it with historical patterns, to identify potential fraud and trigger alerts for further investigation. More importantly, AI and ML play a crucial role in detecting **zero-day attacks** by identifying suspicious patterns and behaviours that may indicate the presence of a previously unknown threat.

Investigate: AI also plays a critical role in investigating current and potential security threats. AI algorithms can be used to analyse data from diverse sources, including social media, forums and threat intelligence feeds, to identify emerging threats and gather valuable insights. AI-powered tools can also scan networks and systems to **identify vulnerabilities or weak points** in the security infrastructure, to identify areas that require immediate attention or remediation. It is also an effective way to automate the collection and analysis of digital evidence, which aids in determining the **root cause of the breach** quickly and effectively. For example, in a complex global supply chain, AI-powered tools can continuously monitor various network endpoints, such as supplier communication systems, inventory databases and logistics management platforms. When an anomaly is detected, AI algorithms can flag it as a potential vulnerability or security weakness.

Respond: AI can also play a critical role in minimizing the impact of threats by **automating security incidents response** effectively. AI algorithms can analyse large volumes of data in real-time enabling organizations to reduce the time it takes to respond to potential security breaches. For instance, security operations centres

(SOCs) across industries leverage AI to identify security incidents more quickly and accurately and automate incident response actions such as isolating infected devices, blocking malicious IP addresses, quarantining suspicious files, etc., which would minimize the time and effort required to contain the incident.

As you can see, AI has the capability to dramatically enhance the data security efforts for an organization. During my time in Flipkart, we had about 150–200 experts, a team of some of the best data scientists in the country, working as part of the 'trust and safety team', whose sole job was to work round the clock to monitor and prevent fraudulent transactions. However, we were always playing catch up. By the time our team could figure out a solution to a potential fraud and deploy it, newer threats would have already emerged, making it very difficult for even the sharpest of minds to keep pace. Now with AI-enabled solutions organizations will hopefully have the opportunity to be more proactive and not reactive to the ever-evolving security threats and threat actors.

However, AI can also increase data security risks

While AI can be a hero of the data security story, the hero does come with a dark side! As we are moving into the AI age, the data security threats are constantly evolving at an even faster pace. Not only that, the complexity and scope of these threats will continue to become bigger and bigger. AI can also be leveraged for more sophisticated attacks, posing a deeper and much more complex security threat. And while AI can help address these issues to a large extent, it can also be exploited for malicious purposes, making it that much more difficult to safeguard against potential risks. For example, password cracking or the brute force attack, using AI is becoming a growing concern. A recent experiment with an AI-powered password cracking tool called PassGAN, done on 15 million commonly used passwords, revealed that 51 per cent of them could be cracked in just one minute, 71 per cent in a day and 81 per cent in just a month.⁹ Now imagine the potential disruption that can be

caused by AI attacks like Vishing (voice cloning appearing to be someone legit), or DeepFakes (where pictures, videos are used for deception). For example, a political leader's DeepFake making a racist statement can trigger riots.

An even more significant concern arises as organizations prepare to implement and scale AI, which can potentially expose them to greater risks. This is a consequence of AI systems becoming increasingly intricate and interconnected, often relying on open and interconnected data sets, increasing the vulnerability in terms of more access points. As a result, these AI systems or models that rely on a continuous feedback loop of data to learn, adapt and improve their performance over time are more susceptible to malicious attacks or manipulation, especially in the case of the Gen AI models that are getting leveraged more and more by organizations today.

This is why organizations are increasingly concerned about the potential leakage, theft or misuse through these models for data generation and manipulation—a topic which is already emerging as a major concern in the client conversations we are having today at my firm Incedo. Gen AI models, trained on broad data sets encompassing text, code and images, pose two key security risks. Firstly, there's the danger of accidental data leakage, where proprietary information may unintentionally become part of a large language model, leading to gradual leaks. For example, let's say if employees start putting company confidential codes for AI to review, it will store those codes as part of terms and conditions and can reproduce it to competitors. Secondly, there's the risk of data poisoning, involving the corruption of proprietary data, resulting in the model producing inaccurate outputs. Additionally, there is a growing fear that malicious actors could employ Gen AI to create convincing fake documents, emails or other content that may contain sensitive data, leading to privacy breaches, identity theft, or misinformation campaigns.

Although this technology is pretty nascent right now, we have already started seeing evidence of misuse. Case in point is the 2019 incident, when a UK-based energy firm fell victim to a Deepfake-based scam resulting in the fraudulent extraction of €2,20,000

(\$2,43,000). During a phone call, the perpetrator impersonated the CEO of the company's German parent corporation. The deception was so convincing that the CEO of the UK subsidiary had no doubt he was speaking with his boss, as the perpetrator was able to imitate the boss's German accent, tone and speech mannerisms flawlessly.^{[10](#)}

So, while AI is a crucial component to solving the data security issue, it is also responsible for creating new issues in ways that we cannot fully comprehend at this point and therefore require careful consideration. The complexity of AI systems, the implications of adopting Gen AI, potential vulnerabilities, adversarial attacks and ethical implications, require ongoing research, robust security practices and a proactive approach to address these emerging concerns effectively.

Cybersecurity, once relegated to a somewhat limited and relatively inconspicuous corner of the IT landscape, has now taken centre stage and gone mainstream. In the not-so-distant past, it was often considered a specialized concern, handled by a select group of experts in the shadows. Today it is one of the fastest growing domains. The cybersecurity market has already reached \$172.32 billion in 2023, and is expected to reach \$424.97 billion in 2030, growing at a CAGR of 13.8 per cent.^{[11](#)} It has undoubtedly become a fundamental concern in the digital age, impacting everything from personal privacy to national security and corporate survival.

Key takeaways

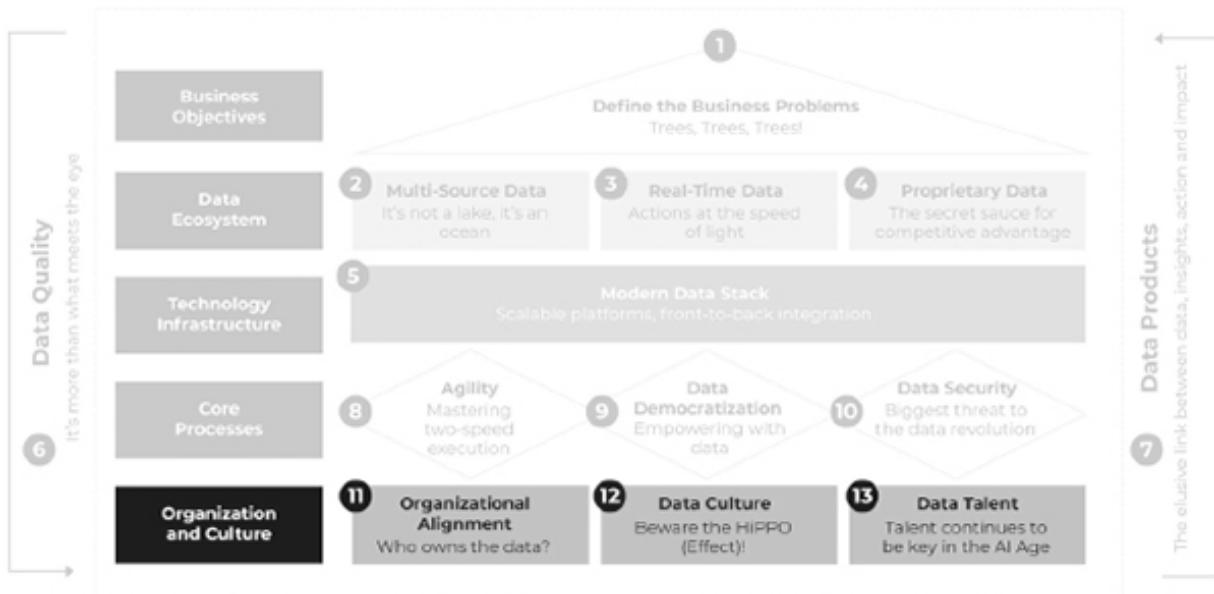
- The exponential growth of data and evolving cyber threats make robust data security vital for the survival and integrity of organizations in the digital age.
- Organizations are rapidly shifting from the traditional 'castle and moat' network security model to a 'zero-trust' security framework, which prioritizes continuous authentication and cautious access verification for all

individuals, devices, or applications seeking network access and is tailored to specific persona or role.

- Data security versus data democratization is one of the key paradoxes of the data world, underscoring the inherent tension between balancing robust data protection with the imperative of enabling widespread data access and utilization within an organization.
- A context-driven approach, keeping in mind three major considerations—data criticality, user persona and role, and activity patterns—is necessary for building an effective zero-trust framework, ensuring a balance between security and access.
- There are four critical levers that come into play for effective execution of a context-driven zero-trust security framework—a well-connected network of technology solutions, robust processes, effective governance, risk and compliance mechanisms and a strong cybersecurity culture.
- AI is a pivotal force in data security across the entire value chain, from adaptive authentication and threat detection to swift incident investigation and automated response, enhancing protection against evolving cyber threats.

LAYER 5

ORGANIZATION AND CULTURE



'Soft issues that are key to unlock potential'

Organizational Alignment

Who Owns the Data?

'The single word that matters most, I think, to keep the company productive as it grows is alignment.'

—Sam Altman,
CEO, OpenAI

Now that the core processes have been addressed and established, let's move on to the next layer, 'organization and culture', which largely focuses on the people aspect of the USF framework. This layer addresses the skills, mindsets and behaviours, and the organization DNA required to drive transformational value from data. This layer talks about how best to align people, how to build a data-driven culture and what are the key skills to focus on in building the data talent pool.

One of the key issues which we talked about in detail in Chapter 5, The Root Cause, is the organizational silos that most organizations struggle with, limiting their ability to generate insights which are relevant for the business, jointly defining and solving the business problems and driving actions from the insights generated. So, in this chapter, I will focus on the first component of this layer, the eleventh component of the USF, which is organizational alignment—what is the right way to align people in the organization to be most effective with data. Because unless you do that, organizations will fail to address the silos that hinder collaboration between various teams, which is critical to unleashing the full potential of data.

As we will see in this chapter, this is not an easy problem to solve because there are a number of considerations that go into building

effective organizational alignment. It is not a one size fits all solution. It requires organizations to make strategic choices to figure out the model that works well for them.

What is organizational alignment?

Organizational alignment refers to the degree to which individuals and processes within an organization are integrated, directed and actively working towards the achievement of the organization's strategic objectives. In simple terms, it means the ability of an organization to get everyone on the same page on what needs to be done and how to do it.

Organizational alignment is a critical enabler that drives the effectiveness and success of data initiatives within organizations. A well aligned organization operates like a well-oiled machine where all the components work harmoniously together to achieve its goals and objectives smoothly and efficiently. When all aspects of the organization are aligned towards a common goal, it fosters a cohesive and unified approach to leveraging data to its full potential. An effective alignment hinges on three key considerations—organizational structure, decision-making and execution. And when the organization is well-aligned on these three, it enables them to drive higher impact from their data initiatives.

The traditional approach and its pitfalls

Traditionally, organizations have predominantly followed a functional alignment approach where teams, such as business, IT and operations, have been responsible for specific functions within the organization. The business team focuses on the strategic and operational aspects of the organization, shaping the direction and goals. The IT team is responsible for managing and maintaining the technology infrastructure, systems and applications. The operations team handles the day-to-day processes and ensures efficient execution.

Under the traditional approach, decision-making is often centralized at the leadership level, where managers and leaders make decisions based on their domain expertise and their understanding of the business requirements. This centralized decision-making can provide consistency but may result in a long-drawn decision-making process.

The functional alignment also leads to these teams being considered as the natural owners of the data generated or utilized within their respective functions. And as the natural owners of data within their functions, each team has distinct requirements and use cases for data. This data ownership within each function can create data silos, limiting access to data across the organization and hindering the ability to leverage data holistically for strategic decision-making. A study revealed that half of the companies that saw more than 10 per cent revenue growth have their marketing and IT teams work together on a shared vision. In contrast more than half of those with the lowest rates of revenue growth admitted that their CMO and CTO rarely interacted, which impacted their ability to create and deliver omnichannel customer experiences.¹

And why is this not ideal in the Big Data world?

In the Big Data world, where organizations operate as a complex ecosystem of technology, networks and people, there are multiple stakeholders generating and consuming data across the entire value chain of the organization. This complex scenario makes it difficult to assign clear ownership or accountability of data to a single person or team, raising an important question—‘**Who owns the data?**’ This lack of clear ownership can result in data-quality issues anywhere across the complex data ecosystem.

Traditional organizational structures, mindsets and working methods can exacerbate these challenges. The siloed working of business, IT and operations teams prevents the joint definition of business problems and goals from the beginning. And when teams work in isolation, without input and collaboration from other stakeholders, the generated insights can be quite off the mark or

difficult to implement. The business team may lack technical expertise in handling and analysing data, leading to sub-optimal insights. On the other hand, the IT and operations teams may lack a deep understanding of the end customer and business context, resulting in solutions that do not align with the organization's objectives. This lack of collaboration and joint problem definition at the outset hampers the ability to determine the right metrics, identify the right data sources and develop effective data-driven solutions, impacting the overall effectiveness of any data initiative. According to a study, 79 per cent of decision-makers find cross-functional collaboration challenging due to the fact that collaboration workflows are frequently conducted in isolated or separate silos.²

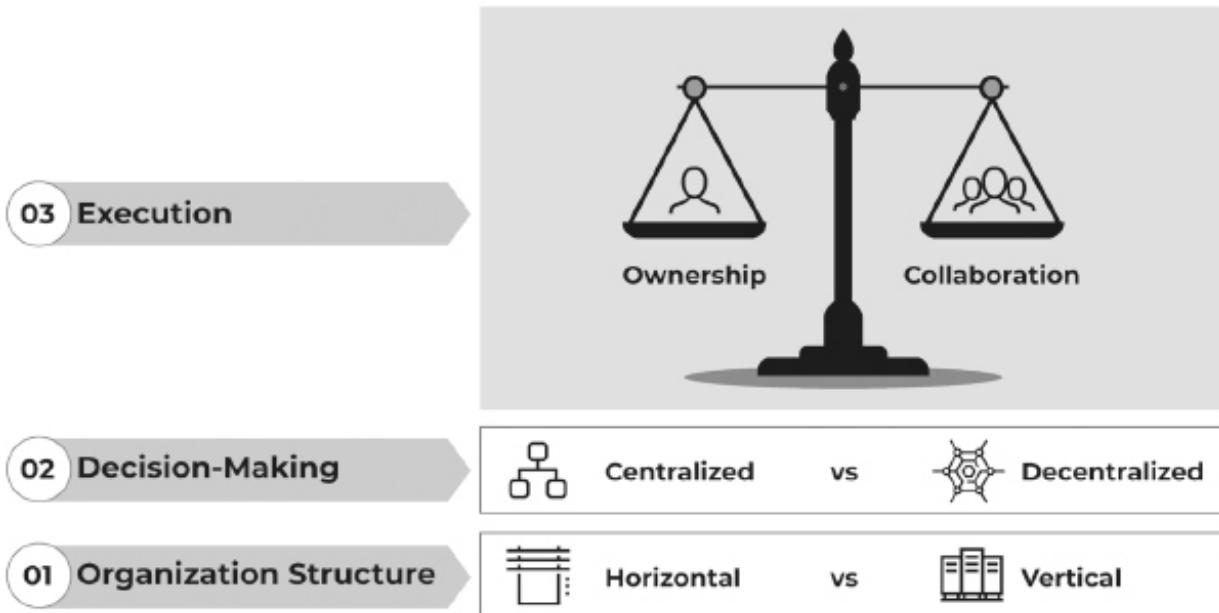
So how do we create an approach that enables organizations to build organizational alignment that keeps data at its core? There are three key considerations that must be taken into account to build effective alignment in an organization.

The three key considerations for organizational alignment

For the success of any data initiative, organizations need to align three aspects: **the structure**—design and arrangement of departments and teams; **the decision-making processes**—the processes and authority around making decisions, and **the execution**—defining clear ownership and ensuring accountability while fostering collaboration.

How organizations structure their teams and the extent of decentralization of decision-making are important choices that organizations need to make based on the context of the business, the industry dynamics, organization maturity and their strategic objectives. The third aspect, effective execution requires bringing a balance between ownership and collaboration. Let me elaborate on each of these three key considerations.

Organizational alignment is a combination of three key considerations



Organizational structure: How will the people engage?

Organizational structure refers to the way teams are organized or arranged—departments, teams and reporting relationships within an organization. It involves designing a structure that enables efficient coordination and information flow within and between different departments and teams so that they can work together to achieve common goals and objectives. There are predominantly two ways organizations can structure.

A **horizontal organization structure** entails capability-based alignment where organizations form teams based on their specific functional capabilities. Employees with complementary skills and expertise work together to support multiple products and stakeholders. This structure is ideal to build efficiency through reusability across multiple products, deliver consistent and standardized offerings, and develop depth of expertise to deliver high-quality solutions. It also allows for more engagement within the capability layer, fostering innovation. For example, the data and analytics function when structured as a horizontal capability, can provide expertise that spans across diverse industry and client

groups in the organization. This horizontal approach centralizes the data and analytics expertise and enables the organization to develop depth of specialized technical knowledge and insights, which can be leveraged across the organization. The depth of knowledge and insights helps develop best practices, foster innovation, build new capabilities (for example, core models, data products) and promote their reusability across multiple areas of the organization.

On the other hand, **vertical structures** emphasize a more client-centric approach where each team works independently to deliver end-to-end on a goal or objective for a client or set of clients (combined into an industry vertical). It brings diverse teams together, fostering a strong sense of purpose and clarity of objectives. This structure enables them to be more agile and responsive. Additionally, this structure enables end-to-end ownership and customer-centricity which drives better outcomes. This structure helps in building deeper industry and domain competencies across the stack.

For example, in a vertical organizational structure, data and analytics teams will be aligned to serve specific industry or customer segments, tailoring their focus to their specific needs. Each customer segment team will have a range of functional capabilities housed within that team. As part of these functional capabilities, the data and analytics team will focus on building in-depth domain knowledge and capabilities to understand the pain points and create solutions for that specific industry or customer segment.

While both horizontal and vertical structures have their advantages, each one has its own inherent challenges. Horizontal alignment often can lack customer-centricity and outcome focus. As a result, measuring business value and impact becomes difficult, ultimately resulting in reduced flexibility and speed in achieving desired results. Horizontal alignment can lead to more of an internal orientation and also requires immense coordination, especially in a large organization. Teams can become too focused on their capabilities and lose sight of the broader organizational goals. On the other hand, vertical alignment may result in silos leading to reduced knowledge-sharing and poor collaboration between teams.

Different vertical teams might unknowingly duplicate efforts or work on similar tasks independently, leading to inefficiencies and wasted resources. It also becomes difficult to build and scale up organization-wide capabilities.

Clearly, there is no one right answer between horizontal alignment and vertical alignment. Organizations have to make a choice between them or a hybrid, depending on the context of their business and organizational maturity.

Decision-making: Who will take the call?

Organizations have to define the process by which decisions are made and the distribution of decision-making authority within the organization. It entails establishing whether decision-making is centralized or decentralized, defining decision rights and responsibilities, and ensuring that decision-making practices are consistent across the organization. **Centralized decision-making** processes enable the organization to scale operations efficiently and maintain consistency in their delivery processes. By consolidating decision-making authority at higher levels, they can efficiently coordinate resources, allocate tasks and streamline processes. For example, a small retail chain with multiple stores will likely have a centralized decision-making structure. Major decisions, such as pricing strategies, product selection and marketing campaigns, would be made by the company's executive team at the headquarters. Store managers and employees at individual locations would likely have little input in these decisions and would be expected to follow the guidelines set by the central authority.

In contrast, in a **decentralized decision-making** approach, decision-making authority is distributed across various levels or teams within the organization. This allows for greater flexibility and adaptability in responding to business requirements quickly. It also promotes collaboration and ownership at different levels. For example, a large multinational retail chain that operates a chain of hypermarkets, discount department stores and grocery stores, would likely have more decentralized decision-making. Depending on the

extent of decentralization, at the region or store level, either the regional teams or the store managers would have a level of autonomy in decision-making for that region or store. Decisions around product pricing, marketing campaign, or product assortments would be made depending on the regional demand, market conditions, customer preferences, cultural preferences and even regulatory requirements.

In terms of disadvantages, centralized decision-making slows down the entire process, becoming a bottleneck in achieving speed and agility in making decisions or driving action. Decentralization, on the other hand can result in silos, inconsistencies, limiting collaboration and result in duplication of efforts across various teams.

In practice, many organizations adopt a hybrid approach, combining elements of both centralized and decentralized decision-making to suit their unique needs and organizational culture. The extent of decentralization can vary based on the size of the organization, the nature of the decisions being made and the organization's need for control, speed and adaptability.

Execution: How will it be done?

Execution alignment defines how work is carried out. It determines who owns the initiative and designates responsibilities across team members to deliver the outcome. There are predominantly two aspects that organizations have to focus on here. **Ownership** which emphasizes clarity in roles and responsibilities. Each team or department takes ownership of specific aspects of the initiative, ensuring that their assigned tasks and objectives are fulfilled.

Collaboration focuses on fostering cooperation and teamwork across different teams or departments. It encourages cross-functional coordination and shared responsibilities, leading to a holistic and integrated approach towards achieving the objectives.

The inherent tension between ownership and collaboration

Historically, organizations have often focused more on ownership than on collaboration, where clear ownership and responsibility were assigned to specific individuals or teams. While this approach may have worked well in an earlier era, the rise of the digital age and the data-first world has introduced greater complexities. In this era, organizations operate as a complex value chain and solving most problems requires collaboration across multiple stakeholders, who bring in varied expertise—domain, technical, data, business, etc. Also, the data required to solve any problem is often elaborate and needs to come from diverse data sources, not confined to one team or function. This interdependence of different teams heightens the need for cross-functional collaboration. However, in such a scenario, where multiple stakeholders are involved, and the data being leveraged comes from various interconnected systems and networks, it becomes extremely difficult to define clear (and, therefore, narrow) ownership and accountability for any data initiative. As a result, this inherent tension between ownership and collaboration is becoming increasingly important for organizations to address.

A data product-centric approach—The balancing act!

Data product-centric approach to organizational alignment is an approach that places data products at the forefront of how teams are structured. It involves aligning teams, resources and workflows around the development, deployment, continuous monitoring and improvement of data products. In this alignment approach, the organization structures itself around delivering outcomes through data products. Managing and delivering data products effectively requires a collaborative and cross-functional approach, which is where the concept of ‘pods’ comes into play. Pods are cross-functional teams that bring together individuals with capabilities across various layers of the data stack. And these pods are effective in addressing the inherent tension between ownership and collaboration in the following ways:

1. **Clear objectives:** The data product has well-defined objectives that align with the overall business goals. These objectives are the driving force behind the data product's development and evolution which naturally ensures ownership and accountability.
2. **Outcome-focused collaboration:** Since cross-functional teams come together to deliver on the outcome, data products naturally enable collaboration. The diverse skills within the pod collaborate closely to develop and deliver the data product to achieve a specific outcome.
3. **Accountability for outcomes:** Teams working on the data product share accountability for its success. At the same time each team member is accountable for their specific layer of the data product. They collectively take ownership of meeting the product's objectives and are jointly responsible for its performance and impact.
4. **Continuous development:** Since the product team is the owner of the product, they are also responsible for incorporating continuous improvements based on user feedback, data insights, changing business requirements and adopting best practices from other data product teams.
5. **Data governance and quality:** The product team is also the owner of the underlying data and are therefore responsible to maintain data quality, security and compliance throughout the data product's life cycle. This ensures that the underlying data is updated and ready to use at all times.
6. **Autonomy in decision-making:** The product managers who own the products have a level of autonomy in decision-making, enabling them to make faster and more informed choices anchored in customer-centricity.

As you can see, this approach emphasizes driving focus and accountability across all capabilities required to deliver a specific data product's objective. It aligns with the growing importance of data-driven decision-making and the increasing complexity of data initiatives in the modern business landscape.

Spotify successfully achieves a product-centric organizational alignment³

The Swedish company Spotify offers a good example of how organizations can strategically structure their operations and decision-making to thrive in the ever-changing data-first world. The Spotify model expertly balances autonomy, collaboration, accountability and quality. By centring its operations around products, Spotify demonstrates how businesses can achieve agility and foster cohesive alignment across the entire organization.

Spotify restructured itself and identified key elements to guide its people and teams. Now, let us delve into these elements, known as 'squads, tribes, chapters and guilds', to understand how this works.

Spotify's employees are organized into cross-functional teams, known as **squads**, which has been a fundamental way of organizing teams for years now. Each squad typically consists of fewer than eight members who hold end-to-end responsibility and work collaboratively towards delivering a product feature and drive the desired business outcome. They have complete visibility and accountability for the successes and failures of their features throughout the product's lifecycle. Empowered with decision-making authority over what to build, how to build it and who to collaborate with, squads embody autonomy and ownership. This approach fosters swift decision-making and empowers squads to assume complete accountability for product delivery. Balancing autonomy and accountability, squads maintain a strong customer focus on achieving outcomes with speed and innovation.

Now, when multiple squads collaborate on the same feature area which has multiple features, they form a **tribe**, creating alignment

and cross-collaboration. Tribes serve as a platform for coordination, consisting of 40–150 individuals to ensure alignment and productivity. Rather than enforcing standardization, Spotify values cross-pollination. If enough squads adopt a particular tool and find success, it becomes the de facto standard, promoting collaboration. Each tribe is led by a dedicated leader who fosters a productive and innovative environment for the squads. This setup promotes strong alignment on business outcome, enables end-to-end ownership and strong customer focus, while encouraging collaboration between squads.

The **chapters** are an essential element in Spotify's model, fostering stronger capability-building and innovation. These chapters connect tribes through specific competencies like quality assistance, agile coaching and web development, facilitating smooth exchange of expertise and knowledge. This approach allows for the establishment of common standards and best practices, harnessing the advantages of scale while encouraging a culture of innovation. With formal managers providing coaching and mentorship, chapters ensure the development of specialized skills within each competency area. Squad members, in addition to their tribal affiliations, belong to chapters where they collaborate with others who share similar skill sets, creating a dynamic space for knowledge sharing and expertise exchange. This unique interplay of chapters contributes to Spotify's continual growth and innovation.

As we delve deeper into Spotify's organizational structure, we encounter **guilds**—the lightweight communities that foster knowledge sharing across chapters and squads, covering various areas from leadership to continuous delivery. Guilds unite experts, facilitating the exchange of insights, tools and best practices, driving collaboration beyond specific competencies. This dynamic interplay of guilds contributes to Spotify's culture of innovation and continual growth.

Furthermore, Spotify recognizes the importance of alignment of different perspectives when working on various features of a product. The **trio**, consisting of a tribe lead, product lead and design lead, ensures continuous alignment when working on feature areas

within a tribe. By bringing together these key roles, Spotify aims to mitigate dependency risks among squads and promote alignment of key decisions for improved efficiency and consistency.

While the names of these teams may have evolved, as I write this, in essence the structure remains quite similar. The Spotify model has gained widespread popularity as a source of inspiration for organizations seeking to cultivate alignment, autonomy and a culture of continuous improvement. This model places significant emphasis on quick decision-making and self-organization, empowering teams to adapt swiftly to ever-changing circumstances. Its flexible structure, characterized by squads, tribes, chapters and guilds, offers scalability and versatility, allowing companies to tailor their approach of decision-making and team structure based on specific product requirements. One of the model's key strengths is its ability to strike a harmonious balance between individual ownership and collaborative efforts. By fostering a work environment that encourages innovation and independence, while still aligning with broader organizational goals, it promotes a sense of ownership and accountability among team members.

Notably, Spotify's willingness to continuously evolve and adapt its model demonstrates a learning-oriented mindset that resonates with organizations seeking to thrive in today's dynamic business landscape. While the Spotify model may not be a one-size-fits-all solution for every organization, its core principles hold valuable insights for any company striving to create a high-performing and inspiring work environment. In fact, many of our clients at Incedo, are working towards implementing a Spotify-like model to achieve better alignment. Encouraging businesses to make explicit choices in their operating model and working culture, the model emphasizes the importance of aligning all elements to nurture a dynamic and engaging atmosphere that empowers teams to flourish and drive transformative results.

Key takeaways

- Organizational alignment refers to the degree to which individuals and processes within an organization are integrated, directed and actively working towards the achievement of the organization's strategic objectives.
- For the success of any data initiative, organizations need to effectively synchronize **the structure**—design and arrangement of departments and teams; **the decision-making processes**—the authority and methods for making decisions; and **the execution**—defining clear ownership and ensuring accountability while fostering collaboration.
- Horizontal alignment boosts efficiency and expertise sharing but may lack outcome focus, while vertical alignment prioritizes customer-centricity and agility but can result in silos and inefficiencies. Choosing the right alignment depends on an organization's goals, industry dynamics and organizational maturity.
- Centralized decision-making processes are effective for maintaining consistency and scalability in operations, while decentralized approaches offer speed and agility in decision-making or driving actions. Organizations must adopt a hybrid approach as per their unique business objectives and culture.
- For effective execution, organizations should consider adopting a data product-centric approach. This would help address the inherent tension between ownership and collaboration by aligning teams, with varied expertise, and clearly assigning end-to-end ownership of the product and the underlying data.

18

Data Culture

Beware of the HiPPO Effect!

'If we have data, let's look at data. If all we have are opinions, let's go with mine.'

—Jim Barksdale,
Ex-president and CEO, Netscape

As the saying goes, 'You can lead a horse to water, but you can't make it drink'. All the effort and resources you invest in working on the USF components that I have talked about so far in the book, won't be effective in the long run without a supportive and enabling data culture. Despite such abundance of data all around, most organizations still follow the traditional way of working—the HiPPO (highest paid person's opinion) culture—where the opinion of the highest paid person takes precedence over everything else.

In this chapter, I highlight the importance of data culture and in particular the dangers of the HiPPO effect. I aim to equip you with the strategies needed to counter the HiPPO effect and build a data-driven culture.

What is data culture?

Corporate culture is defined by the *Cambridge Business English Dictionary* as 'the beliefs and ideas that a company has and the way in which they affect how it does business and how its employees behave.' In a simpler way, I define corporate culture as 'what you do when nobody's looking!'

If an organization has a strong ‘belief’ in being data-driven, it should manifest both in the organization’s processes and actions of employees—their behaviours. So, data culture is ‘the set of attitudes, practices and behaviours that encourages leveraging and maximizing the use of data in decision-making processes across the organization’. The litmus test for a data culture is when data becomes an integral part of the decision-making process at all levels across the organization, which requires employees at all levels to be empowered and trained to effectively use data and draw insights to make informed decisions.

Let’s revisit the third symptom of the Data Paradox, discussed in Chapter 4, The Data Paradox, ‘Teams are not doing much with the data they have.’ There are three key reasons why this is the case. First, that many of the business users lack the technical background needed to understand and utilize available data. Also lack of joint problem-solving, especially when identifying and defining problems, results in lower confidence and distrust in using data for decision-making. Second, the absence of structured processes for data sharing and data-driven decision-making creates a negative cycle, leading to ineffective use of data across the organization. And third, leaders who do not encourage a data-driven culture and lead by example become a roadblock to its adoption within the organization. Most traditional organizations are bogged down by these challenges. This is why, only 39 per cent of the respondents in a 2019 survey said that their company possesses a robust cultural commitment to utilizing data-driven insights for decision-making.¹

This underscores the urgency of the need to initiate a fundamental shift in the traditional approach, to a culture where data is valued and utilized effectively across all levels of the organization. However, a significant obstacle in establishing a data-driven culture is the HiPPO effect.

The HiPPO looms large

In the vast savannah of the business world fuelled with Big Data, a phenomenon holds significant weight—the HiPPO, the opinion of a person whose job title and related salary put them in a position of influence. This confident and towering figure casts a shadow over the decision-making landscape, armed with experience, authority and personal beliefs. Although relying on the HiPPO is a very natural approach to decision-making, owing to the leader's experience and authority, there lies a hidden danger of the HiPPO effect.

The HiPPO effect is characterized by a decision-making culture that prioritizes the HiPPO's viewpoint over data-driven insights, evidence and the collective knowledge of the team. The term was first mentioned in the book *Web Analytics: An Hour a Day* way back in 2007, in which the author highlighted how in situations where a difficult decision needs to be made and there is a lack of data or data analysis to provide clear direction, the group often relies on the HiPPO.² But here's the thing, HiPPO isn't some new phenomenon but is intrinsic to many situations. We've all encountered such scenarios, be it at our workplace or in personal lives. You know what I'm talking about, right?

Now, in a data-scarce world, relying on the opinion of one authoritative person might have been the only option, whether it was the village elder, the priest or the business owner. However, over the years as data has grown exponentially, this HiPPO effect can have a crippling impact on the decision-making process. Whether it is an unsaid organizational norm, lack of confidence or fear of going against the authority, it has become a system hard to break out of. According to a study, although 80 per cent of survey participants said they used data in their job responsibilities and 73 per cent used data to inform their decisions, a significant 84 per cent indicated that managerial judgement remains influential in making critical decisions.³ That's quite a contradiction!

The perils of the HiPPO effect

The detrimental outcome of the HiPPO effect has been highlighted by a prominent study contrasting the impact of Asian vs Western cultures on aviation safety. Observations indicate that cultures with a high-power distance—where hierarchies that are clearly defined and therefore unquestioned, such as Asian cultures, experience more crashes compared to cultures with high individualism, like Western ones, which have fewer accidents. Notably, incidents like Korean Air crashes were attributed to communication issues within the cockpit. Cultural norms emphasizing unquestioning respect for authority led junior pilots to avoid reporting safety concerns, fearing damage to their relationships with colleagues and superiors. Interestingly, studies show that airlines from countries with less rigid hierarchies tend to be safer.⁴

Granted, most decisions are not as life-threatening as an air crash, nevertheless, HiPPO culture poses a significant detriment to the entire decision-making process. Decisions based on authority figures' gut instincts or personal beliefs which aren't backed by sufficient data and evidence, can result in biased and flawed decision-making. This inherent flaw has been proven through extensive scientific research on self-awareness—Involving over 800 studies and a survey of thousands of employees. It suggests that although most people think they are self-aware, self-awareness is a rare quality. The survey reveals that while 95 per cent of people think they're self-aware, only 10 to 15 per cent actually are.⁵ And that very few leaders are able to actively focus and work on balancing both internal (how we see ourselves) and external (how others see us) aspects to build self-awareness. The survey also highlighted that in most cases experience and power leads to a false sense of confidence in one's performance and overconfidence in self-knowledge. Only a few highly self-aware leaders frequently seek critical feedback to become more self-aware. And lastly, very few leaders who are truly self-aware ask the 'what' rather than the 'why' to stay objective, future-focused and empowered to act.⁶

So, this HiPPO approach clearly has some significant downside to it:

Lack of diverse perspectives: When decisions are predominantly HiPPO driven, it is often based on one person's version of truth. It undermines the importance of supporting viewpoints, diverse thinking and data-driven evidence within the organization. By exclusively following the HiPPO's lead, valuable information and alternative perspectives are disregarded in the decision-making process.

In my previous book, *Winning in the Digital Age*, I emphasized diversity—be it gender, age or race, geography, etc.—as one of the fundamental requirements for any world-class organization striving for excellence in the digital age. The absence of diverse perspectives hampers the ability to make well rounded, informed decisions, increasing organizational risks. Research has consistently shown that diversity has a significant impact on an organization's financial performance.

In a study involving 1700 employees, six dimensions of diversity, including gender, age, nationality, career path, industry background and education, were examined for correlation between diverse management teams and financial performance. The findings revealed that companies with above-average management diversity produced 19 per cent higher proportion of revenues from innovation (from new products and services) and achieved ~10 per cent higher EBIT (earnings before interest and taxes) margins compared to those with below-average diversity, as teams with diverse members were better at addressing market segments aligning to their teams' demographic composition.⁷

Risk of bias in decision-making: The HiPPO often goes unchallenged, introducing the risk of biases affecting decision-making. These biases can include authority bias (relying heavily on experts' decisions), anchoring bias (giving undue weight to initial information) and homogeneity bias (assuming unanimous agreement). These biases limit the assessment of all options and

impact final decisions. The famous Milgram experiment, conducted by psychologist Stanley Milgram at Yale University in 1963, illustrates the negative consequences of authority bias. This study investigated obedience to authority figures versus following one's conscience. By examining justifications for World War II genocide, the study revealed that many participants blindly followed authority figures' instructions, even when they believed their actions were harmful, shedding light on the power of authority in shaping human behaviour.⁸

Lack of alignment: The HiPPO effect invites resistance to change and a sense of exclusion due to limited involvement in the decision-making process, communication gaps, fear of the unknown and lack of trust. A study of more than 30,000 employees and leaders, investigating the factors influencing employee resistance to change revealed that only 15 per cent of the employees understood the rationale behind every strategic decision of their organization. The study suggests that if an organization encounters substantial resistance to change among employees, often the reason may be the lack of understanding of the rationale behind the proposed changes.⁹

Lack of accountability and ownership: Under the HiPPO effect, teams are not fully involved in the decision-making process, so they do not feel empowered and therefore lack the drive to take ownership of the task assigned to them. This detachment diminishes their sense of ownership and investment in the outcomes. And this becomes especially crucial in the digital age, where problems are dynamic in nature and tasks are often ambiguous and less defined. Therefore, there is limited value of managers who just follow and execute on instructions, versus leaders who take ownership and drive action.

Taming the HiPPO with data

To counter the HiPPO effect, every decision made should be backed by concrete data and analysis that substantiates the claims and demonstrates that the decision is grounded in facts, increasing accountability and fostering a more objective and informed decision-making process.

While it sounds relatively straightforward, the real challenge lies in making it an ongoing and automatic practice for employees. An executive survey focused on understanding the potential impact of Big Data, and its implications for their businesses reveals that just a little over 24 per cent of the Fortune 1000 companies have succeeded in forging a data culture within their firms so far. The report highlights that despite significant investments in Big Data and AI over the past decade, organizations still lack the execution strategy to bring those investments into fruition.¹⁰ These organizations continue to encounter substantial challenges in their journey towards becoming data driven.

The key levers to building a robust data culture

In the Big Data world, where organizations are inundated with massive volumes of information streaming in from various sources, fostering a data-driven culture has become even more critical and pressing for achieving success. The people who work on data transformation initiatives are more than aware that it is often culture, not technology, that deadlocks their efforts.

Establishing a data culture requires fundamental transformation of the organizational DNA, which would enable data-driven decision-making. Companies that cultivate a strong data culture are notably more adept at extracting value from data.

In my experience, **there are three key levers** that companies could use to foster a culture of data within their organization:

I. Data literacy

Data literacy refers to employees' ability to understand and use data effectively. It goes beyond basic data understanding to include essential skills like data interpretation, critical thinking and the ability to describe the use case and resulting value. It also involves understanding data sources, analytical methods, as well as assessing data quality, potential biases and limitations. Ultimately, data literacy enables individuals to work efficiently with data and related tools and technologies, enhancing their overall effectiveness in utilizing and leveraging data within the organization.

As an early advocate for the vast potential of data and analytics, and having worked as a consultant in my career, I had the opportunity to serve clients from different industries, where I observed how each industry harnessed data in unique ways. Then at Flipkart, I personally witnessed the incredible opportunities data presented and was truly amazed by its profound impact. I was amazed by the deep data-driven culture in the organization. In our daily morning 'stand-up' meetings, which had over 100 attendees, a significant amount of data was presented and various KPIs were discussed. It was truly incredible to witness the active engagement of so many individuals, as they absorbed the data and even questioned its validity. Even the youngest team members were so well-versed in working with data that they could challenge the CXOs.

This experience has reaffirmed my strong belief that data literacy is crucial in fostering a data-driven culture within any organization. And I strongly believe, now more than ever, that it needs to be an organization-wide effort and not limited to a handful of data experts and scientists.

There are multiple ways organizations can incorporate data literacy across all levels:

- 1. Learning and upskilling:** Organizations can enhance their employees' ability to comprehend and work with data through structured learning and upskilling programmes, facilitating more confident decision-making

and meaningful discussions with other colleagues. Considering the ubiquitous nature of data, it is crucial to take a tailored approach when developing data literacy initiatives. It's important to note that not everyone needs to become a data scientist; data literacy is a personalized journey.

Airbnb's Data University is a data-literacy programme that empowers employees to make data-informed decisions. With volunteer faculty members teaching courses globally, covering various skill levels, the programme has seen over 400 courses taught with 6000 registrations. The programme offers different levels of courses, ranging from foundational to advanced topics like machine learning and data pipelines. A customized condensed version called 'Data U Intensive' training provides tailored courses to specific business units and large groups. This format has had a significant impact, boosting participants' confidence, engagement and ability to apply data in their work. It has also enhanced data fluency, enabling teams to make informed decisions, create tailored solutions and measure their impact effectively.¹¹

2. **Inspiring data-driven decision-making:** To educate and motivate others about effective data utilization, leaders can incorporate examples and success stories into awareness campaigns. By showcasing real-life scenarios where data has been successfully used to drive significant business impact, leaders can provide concrete illustrations on the benefits of leveraging data effectively. Sharing the outcomes of such initiatives with the entire organization helps individuals witness firsthand how data-driven decision-making can lead to positive outcomes. By using storytelling and real-world examples, data concepts can be made more relatable and

accessible, empowering employees to apply similar approaches in their own work.

PepsiCo encouraged adoption of their data-driven decision-making platform, PepViz, by demonstrating the significant value it can generate for its retailers. PepsiCo team used the Pepviz analytics platform to enhance retailer performance by optimizing shelf space, increasing basket sizes and identifying opportunities for healthier snacks, which resulted in boosted sales and profitability. They effectively communicated the value generated by PepViz both internally and externally, serving as a compelling example to encourage its adoption by retailers and its employees.¹²

3. **Democratization of data:** As highlighted in Chapter 15, Data Democratization, it is increasingly important to ensure that employees have convenient and unrestricted access to data, while equipping them with the necessary skills, tools and understanding to utilize it effectively. By doing so, organizations can facilitate improved decision-making based on comprehensive information and foster a culture of prompt action-taking. To unlock the true potential of data, it's not just about giving people access to it. We need to create a genuine passion for data-driven decision-making throughout the organization. Democratizing data analytics by providing user-friendly platforms is an important step, as it builds familiarity and trust in data and empowers individuals to find solutions without relying heavily on expensive data scientists.

Uber's data platform, the Uber Data Portal, is a centralized hub that enables employees across various departments and roles to utilize data in their decision-making. Originally launched in 2014 for data scientists and engineers, the platform has since evolved to include self-service analytics capabilities for business users. Notably, city operations teams responsible for managing

Uber's transportation network in each market regularly utilize the platform's data and insights to address specific driver and rider issues. The platform serves as the backbone for driving decision-making processes at the operational level within Uber.¹³

4. **Outcome focus:** Data literacy should not merely be a desirable skill; it should be a driver of tangible outcomes. By connecting data literacy to actual projects and integrating it into daily workflows, organizations can achieve higher adoption rates. It is therefore crucial to link data literacy to practical applications that yield real-world outcomes. But measuring the effectiveness of data literacy initiatives can be difficult without proper metrics. By defining success metrics, organizations can ensure they achieve their desired results from data literacy initiatives.

One of my clients initiated a sales transformation programme to enhance their sales team's effectiveness and boost revenue. They identified a lack of data utilization among the sales team and introduced a data literacy programme with the goal of improving decision-making using data. The programme included modules to enhance the team's ability to use self-service data tools, analyse data for insights, facilitate data-driven communication within teams, and make more informed decisions. As a result, the programme, which was specifically anchored on sales as a key metric, was able to deliver around a 10 per cent increase on that metric.

5. **Data-driven storytelling:** Data literacy encompasses more than just analysis. It also involves the ability to effectively communicate it to others in a way that is easy to understand. Just as literate individuals possess the skills of both reading and writing and can convey a message, data-literate individuals should be proficient in both analysing data and conveying its insights to others

through infographics, visuals or narrative or combination of all. Effective communication about data is an integral aspect of data literacy, enabling individuals to articulate and share the significance of data in a meaningful way. This can be achieved through various channels which could be as big as town hall meetings or regular internal briefing newsletters.

In my opinion, Amazon stands out as a good example of promoting data-driven storytelling through concise six-page narrative memos, which replace slide presentations in their meetings. These documents drive informed discussions through well-crafted narratives, promoting focused discussion and understanding of essential information, unlike lengthy PowerPoint presentations. Moreover, these documents serve as self-contained knowledge repositories, facilitating knowledge transfer across different groups. Amazon's dedication to data-driven storytelling fosters a culture of informed decision-making and collaboration. I will elaborate on it in the next section where I talk about processes.¹⁴

II. Processes

Another lever critical to ensure a robust data culture is establishing the right processes that promote effective and efficient use of data. These processes form the foundation of a strong data culture, by aligning individuals and teams toward the shared objective of using data for positive outcomes. They provide a systematic approach to accessing, analysing and interpreting data, enabling employees to become more comfortable with data, and capable of making informed decisions that drive growth and innovation. Data-driven organizations depend on structured processes to unleash the full potential of data and deliver exceptional products or services to customers.

This is how Amazon does it^{[15](#)}

Amazon has gained widespread recognition for its robust and firmly established data culture. The company has integrated data into its decision-making processes and operations, utilizing it to fuel innovation, enhance customer experiences and elevate overall business performance.

Amazon's decision-making approach, as outlined in the book *Working Backwards: Insights, Stories, and Secrets from Inside Amazon*, is characterized by well-defined, high-quality and rapid decision-making processes, anchored on data, that are consistently applied across the organization. Whether it is at business or functional level, or at organizational strategy level (even when something is to be presented to Jeff Bezos), they follow a consistent format that ensures data is at the core of the decision-making process. At Amazon decision-making is categorized into two types:

The high impact, irreversible, 'one-way door' decisions:

These decisions, also called Type 1, are high-impact and often irreversible. They require meticulous consideration, careful evaluation and consultation. Ideally, individuals should allocate up to 10 per cent of their work week to these decisions. While they can be draining and time-consuming, they demand focused attention. It's important not to make these decisions when experiencing negative emotions or fatigue. For instance, quitting a job on a Monday morning due to temporary despondency can fall into the one-way door decision category. Another example, specific to the business would be a decision around price reduction of a product, which once taken, can be difficult to roll back.

Amazon template for data-driven decision-making

Purpose

<Reason for writing the document and overall direction of the document>

Background

<Context and relevant information about the problem and opportunity anchored in data>

Tenets

<North Star for the proposed solution, which act as the guiding principles (linked to the Amazon Leadership Principles)>

State of the Business

<Insights pertaining to the current state of business under consideration anchored in data>

Lessons Learnt

<Insights, observations and takeaways from past experiences, similar projects, or experiments related to the proposed initiative>

Plan and Goals

<Outline the strategic priorities, plan of action with clear input and output KPIs to track and measure success>

The flexible, reversible, 'two-way door' decisions: These decisions, also known as Type 2, allow for more flexibility, are reversible. Typically, these decisions can be rectified by revisiting the options. So, these can and should be made swiftly, either by individuals with strong judgement or small groups. They can be batched, delegated to team members, or even outsourced to contractors. An example here could be a promotional offering that Amazon rolls out, which, if it fails to do well, can be revisited, iterated upon and rolled out again. To understand Amazon's data-driven decision-making, let's look at their processes. Typically, for 'one-way door' decisions, they use six-page narratives, which provide a structured framework for data-driven decision-making.

The six-page narratives at Amazon follow a standardized structure used across the organization, although individual groups may add their unique touches. These documents aim for a narrative style without bullet points, graphics or excessive details, making them self-explanatory for all readers. They focus on clear objectives, data-driven justifications and defining business outcomes and KPIs. Longer narratives, beyond six pages, contain extra context in an appendix, keeping the narrative's flow intact for easy reference.

Purpose: Every document must start with a purpose, stating the reason for writing the document and declaring the overall direction that the document will take right at the beginning. It is essential to establish a precise scope for the material and convey the intended path that the document will follow.

Background: This section is crucial to provide the context and relevant information about the problem or opportunity the document aims to address in the proposed project. It also includes the data and metrics to quantify the problem. For example, our overall apparel market share is 42 per cent but in women's apparel it is only 18 per cent.

Tenets: Now, tenets are unique to Amazon, and it would not be an exaggeration to say that Amazon is pretty particular about it. In the Amazon culture, it is critical to have a clearly defined North Star for every action taken—a set of foundational guidelines behind any recommendation made. It gives the reader an anchor point from which to evaluate the proposal. These North Stars serve as inspirational pillars from which they can work backwards to figure out the plan to achieve the goal. While the wording may vary, they serve as guiding principles that support and align the overall strategy. An example of a tenet is: we will always adhere to high ethical standards in the way we conduct our business, even if it means forgoing short-term gains for long-term trust and reputation.

State of the business: This section holds significant importance as it aims to provide the reader with hard data on the current state of the business. It is crucial to include detailed information or data in this section as it sets the foundation for the subsequent section, where points of comparison will be presented.

Lessons learned: Amazon places great emphasis on data-driven decision-making. This section typically includes insights, observations and takeaways from past experiences, similar projects or experiments related to the proposed initiative. This section serves to inform decision-makers about the potential risks, challenges and best practices based on historical data and real-world experiences. The goal is to provide a comprehensive snapshot to the reader with the necessary data to understand both the positive and negative aspects of the experience and experimentation so far.

Plans and goals: This section outlines the strategic priorities, usually making up the bulk of the document. It provides detailed plans on how to achieve the goal, execution strategies and ensuring alignment with the goals stated at the beginning of the document. In other words, it serves as the roadmap for achieving the desired outcomes and guides the implementation process.¹⁶ And the key here is to not just focus on outcomes, but to clearly identify the input metrics that would help achieve an outcome or goal, putting emphasis on showing causality. The success of the initiative is measured against these input metrics rather than focusing only on outcomes.

Beyond these sections there is a conclusion and then maybe FAQs or supporting information/documents.

Through the six-page narrative process, Amazon emphasizes the importance of understanding the scope of work, leading to improved outcomes. While crafting these narratives is challenging and requires practice, they serve as a valuable tool for effective decision-making. Every idea or proposal must adhere to this format, promoting focused clarity in communication. Creating these narratives also fosters idea exploration through iterative feedback, forming the basis

of Amazon's robust data-driven culture, driving informed choices that propel the company forward.

III. Role modelling by leaders

Top leaders play a crucial role in promoting a data culture within an organization. Their active engagement with data inspires and influences employees at all levels to prioritize data-driven decision-making. By consistently demonstrating a data-driven approach, leaders earn trust and credibility, establishing a foundation for an organization-wide culture. It also helps break down resistance to change as leaders openly embrace data and communicate its value, showing that it is integral to the organization's success.

This is also backed by the mirroring effect, which suggests that individuals tend to adopt and mirror the behavioural traits they observe in others, particularly those in positions of authority or leadership. In the context of an organization, if a leader actively engages in data analysis, communicates insights based on data, and makes data-informed decisions, employees are more likely to follow suit. Role modelling not only strengthens the organization's data capabilities but also creates a shared understanding and commitment to utilizing data for better decision-making and enhancing data culture. While chief data officers might lead data literacy initiatives, it is crucial for all top executives to actively demonstrate data-driven decision-making.

I believe a role model or leader is someone who truly walks the talk in every aspect. Such role models exhibit four crucial traits that are essential for fostering a strong data-driven culture throughout the organization:

1. **Exemplifying data democratization:** Effective leaders ensure that data is easily accessible to all employees, empowering them with the knowledge and skills to make informed decisions quickly. They value the importance of sharing and are committed to providing regular updates, presenting transparent narratives supported by data at

all times. Their actions serve as a guiding example, inspiring others to follow suit and embrace the same level of transparency and data-driven approach.

2. **Favouring 'data' over 'assertions':** As a leader, I have always emphasized the importance of prioritizing data over assertions, qualitative information, or verbal explanations, often, in my team conversations, referring to it as 'Math over English'. I believe effective leaders demonstrate the ability to set aside assumptions and biases and rely on data to make informed decisions. These leaders carefully consider data when evaluating investment proposals, substantiating opportunities with relevant information, and conducting live tests to observe outcomes.
3. **Tracking metrics for performance:** A leader who values data culture understands that improvement comes from measurement. They ensure that appropriate objectives and metrics are in place for everyone to track and measure their success and growth effectively. These leaders empower their team members to have an outcome mindset which enables them to anchor their programmes on not just output KPIs but clear input KPIs as well.
4. **Embedding data-driven decision-making within values:** Leaders who are truly passionate about data go beyond using data for everyday decisions; they embed a data-driven mindset in their core values and inculcate it in their organization's DNA. These values are not just empty slogans; they are deeply ingrained in their daily actions. For example, in my company Incedo, we have a mission to bridge the gap between the potential and realized value of digital investments and believe effective use of data is fundamental to it, not just in our client work but even internally at every level of decision-

making. Therefore, we have defined one of our core Incedo values to be data-driven in everything we do.^{[17](#)}

In the business world, Jeff Bezos is an embodiment of these key traits—a role model and the passionate leader behind Amazon, who serves as the driving force in encouraging and fostering a data-driven culture that sets the company apart. Every facet of Amazon, including web design, finance and operations, revolves around metrics and measurement. Bezos acknowledges the value of improvement through measurement and instils this belief throughout the organization by establishing new metrics as goals for everyone. What distinguishes Bezos is his willingness to rely on data and challenge existing assumptions, liberating the company from the influence of HiPPOs. A notable instance is when Bezos initially dismissed a proposal for homepage advertising but embraced the idea upon being presented with data, resulting in billions of dollars in revenue through Amazon Advertising.^{[18](#)}

In fact, a key reason Bezos has been successful in building a data-driven culture also stems from one of the sixteen leadership principles at Amazon—have backbone; and disagree and commit—where they hire people, especially in senior roles, who are not afraid to disagree. It is essential that senior leaders do not agree just to fit in or are wary of arguments. Leaders are obliged to respectfully challenge decisions and are not afraid to bring up their point of view, especially against senior leaders, provided it is based on data and hard facts. Because once the decision is made, they have to commit wholly.^{[19](#)}

Such dedication to metrics, data-driven leadership and democratization of data serves as the driving force behind Amazon's success, fostering an environment of innovation and a data-driven culture at Amazon.

Key takeaways

- One of the biggest impediments in enabling a data-driven culture is the HiPPO effect, a decision-making approach that prioritizes the highest paid person's opinion or viewpoint over data-driven insights.
- In the data-abundant world, leaning on the HiPPO may lead to poor or erroneous decision-making, owing to lack of diverse perspectives and heightened risk of bias, and lack of alignment, accountability and ownership amidst the employees.
- The first lever in building a data-driven culture is strong data literacy that can be achieved through continuous learning and upskilling, inspiring through success stories, democratizing data, anchoring literacy programmes on outcomes and encouraging data-driven storytelling.
- The second lever is establishing robust data-driven decision-making processes that enables effective and efficient use of data across all levels in the organization.
- And third critical lever in establishing a successful data culture is the role modelling by top leaders that inspires and encourages employees to become data-driven in their decision-making process.

Data Talent

Talent Continues to Be Key in the AI Age

'Big data isn't about bits, it's about talent.'

—*Douglas Merrill,
Former VP/CIO Engineering, Google*

In the previous chapter on data culture, I talked about the significance of building capabilities across the organization to effectively leverage data, as it serves as the foundation for data-driven decision-making at all levels. However, in order to truly thrive in the data-first world, organizations also need to focus on building the right talent pool that possesses specific data-related skills, expertise and knowledge to unlock the transformational value of data across the data management value chain.

In the Big Data world where the emphasis has shifted towards value realization from data, the role of data talent is expanding beyond traditional skill sets. As a result, data talent must now focus more on delivering value which requires more than traditional data skills. While proficiency in technology or data tools remains crucial, data professionals should also possess deeper domain knowledge as well as business acumen, problem-solving and strong communication skills. Furthermore, traditional data scientist roles will see a significant shift, and the role of data engineer and data architects will become even more important.

In this chapter, I will highlight the key data roles essential to the data-first world, emphasize the increasing significance of acquiring talent with the right skills and address the associated challenges. I

will further explore how the roles and skill sets of data professionals are evolving in the AI age and will continue to do so.

Importance of data talent

In the complex landscape of Big Data, organizations require **specialized talent** to extract value from vast and diverse data sets. This talent excels in data handling, integration, processing and quality assurance. They utilize advanced analytics, ML and statistical models to uncover patterns and generate actionable insights. Additionally, they are proficient in navigating complex data ecosystems, crafting robust data architectures and optimizing data workflows to harness innovative technologies effectively. Professionals skilled in data visualization and storytelling are vital for transforming complex data into compelling narratives, facilitating effective communication and decision-making.

Data talent also serves as a crucial bridge between business, technology and data teams. They facilitate collaboration between stakeholders from both the business and technical sides by aligning business goals, translating business requirements into technical specifications, identifying the data requirements and designing data solutions.

Data talent consists of a variety of roles, not just one

While there is a wide range of roles that make up the specialized data talent pool, here I will focus on a few key roles that I consider to be most critical.

Let me start with the **business analysts**, who translate the business problem into a data problem. They are typically part of business functions and are the go-to people for business leaders for their data or insight needs. They facilitate communication and foster collaboration between business teams and data professionals. Business analysts have a deep understanding of the business situation along with the ability to work with data.

The **data scientists** are skilled professionals who possess expertise in statistics, programming and domain knowledge to extract insights, tackle complex problems and drive data-driven decision-making. They analyse large data sets using techniques like data mining, ML and statistical modelling to uncover patterns and make accurate predictions. Additionally, data scientists work closely with stakeholders to understand business objectives and determine the most suitable data-driven approach to achieve those objectives.

The **data architects** focus on creating an efficient and scalable data infrastructure that supports data storage, processing and analysis. They design data models, schemas and data-integration strategies to ensure data flows smoothly and consistently across different systems and applications. With their expertise in designing and managing data architecture and infrastructure, they define data standards, data-quality guidelines and establish data-security protocols to ensure the integrity, privacy and compliance of data. Data architects collaborate with various stakeholders, including business analysts, data engineers and IT teams to understand data requirements and translate them into robust data architectures.

Organizations typically need to have a variety of specialized data roles



Of course, we cannot overlook the indispensable role of **data engineers** in this intricate ecosystem. They design, build and maintain the systems for data storage, processing and analysis. They design and implement data pipelines to extract, transform and load data from various sources into data storage systems such as databases, data warehouses or data lakes. Data engineers optimize workflows for scalability, performance and data quality. They are skilled in programming languages like Python and SQL, as well as technologies such as Apache Hadoop and Spark. Data engineers play a crucial role in establishing the foundation for data-driven applications and ensuring efficient data-processing systems.

But what good is data without the ability to comprehend and communicate the insights effectively? This is where **data visualization experts** come into play. Armed with a deep understanding of the latest BI tools, they transform complex data insights into visually captivating representations. Through captivating

charts, graphs, interactive graphics, advanced BI tools and increasingly audio/video presentations, they unravel the hidden stories within the data, empowering stakeholders to grasp the significance of the findings at a glance.

Last but not least, the **product managers**, they are the integrators of teams and activities across business data and technology and are end-to-end owners of data products. They collaborate with stakeholders, including data scientists, engineers and business teams, to understand and define the business requirements. They play a key role in driving the solutioning process, defining product requirements, creating an execution roadmap and orchestrating the overall execution of the data product to deliver the business outcome.

However, acquiring high quality, specialized data talent has never been an easy task. Let me explain why.

Acquiring the right talent can be an uphill battle

As I have highlighted in root causes, a key issue of the Big Data world is the lack of alignment between IT, business and operations. To bridge this gap and build data-driven businesses, right talent is required that comes bearing not just data expertise, but also is able to understand the business problem that they need to solve as well as effectively craft compelling stories. While organizations understand the importance of building the right data talent pool, in reality, sourcing the right talent proves to be a formidable challenge.

A survey highlighted that 53 per cent of enterprises identified the scarcity of talent as a major hurdle in deploying Big Data.¹ The shortage in the data field is not limited to core technical and data skills, but it is even more challenging to find data professionals who also possess strong problem-solving abilities and can effectively apply data to address real-world issues.

This challenge stems from two critical factors:

Limited pool of top data talent to choose from

While the total pool of data talent might seem large, in my experience a meagre top 10 per cent of the talent pool really has the necessary skills required to thrive in the world of Big Data. Only this small pool possesses the desired attributes of exceptional problem-solving skills, storytelling and innate business acumen combined with the core data and technical skills. This select group has the extraordinary potential to unlock immense value for organizations, but their exceptional abilities come with a hefty price tag, in comparison to their counterparts. What is even more worrisome is that this already limited pool of top talent is expected to shrink further with the advent of the AI age, which demands professionals who can harness the power of sophisticated AI technologies—to think critically, creatively solve complex problems and leverage AI to drive innovation. It wouldn't be surprising to see the compensation for these skills skyrocketing way beyond current levels.

This further intensifies the complex challenge organizations face in their efforts to attract and retain these highly sought-after individuals. Closing this talent gap remains a daunting task for organizations striving to thrive in today's data-driven landscape. As a result, most companies have to make do with the remaining 90 per cent that possess core data skills but lack distinctive problem-solving abilities and business domain expertise.

Attracting and retaining top data talent is no easy feat

The reality is undeniable: the crème de la crème of data talent who are exceptional problem-solvers tend to favour an open culture that not every organization can provide. Top talent actively pursue organizations that offer an appealing and fulfilling work experience that includes various aspects, like flexible working arrangements, a positive and inclusive workplace culture, opportunities for professional and personal growth, competitive compensation and benefits, as well as a supportive and innovative working environment. But many organizations lack the necessary structure

and processes to provide such flexibility, freedom and culture desired by these exceptional individuals. Additionally, such top talent prefer certain lifestyle standards, which are offered only by certain hubs, such as the Bay Area or Bangalore in India. This inclination often draws them towards successful digital natives like the Googles and Microsofts of the world, which have become magnets for the top 10 per cent of data professionals worldwide because of the culture and opportunities they provide. Traditional organizations face hurdles in building an ecosystem that can attract and retain such top talent. This entails creating an environment that nurtures and supports the growth of these exceptional individuals, providing them with the right culture, abundant opportunities for development and learning, and access to cutting-edge technology and innovation tools.

The talent dilemma we face is nothing new. It has persisted for quite some time, and I can personally relate to it from my experience during my time at McKinsey.

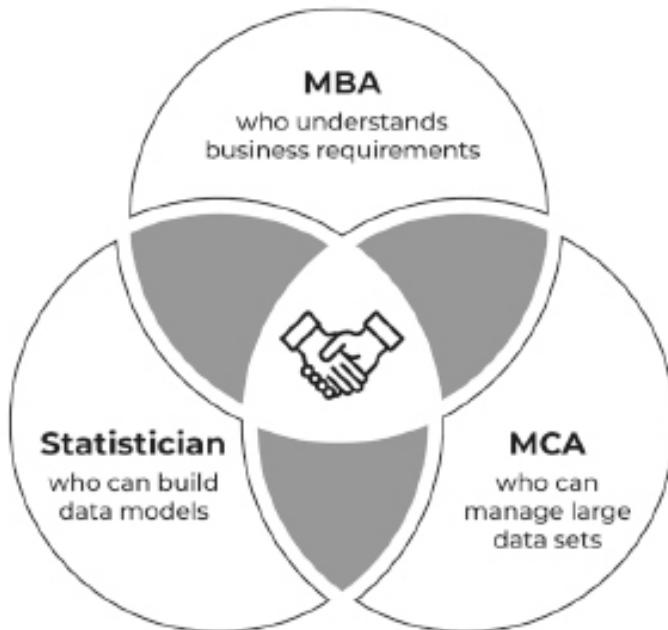
This is how I solved the data talent problem

During my time leading the McKinsey Knowledge Centre (McKC), I was faced with a similar situation which required a creative approach to solve it. At that time, McKinsey had a niche analytics group comprising about ten professionals across New York, London and Dusseldorf. These experts possessed a unique blend of deep statistical knowledge, domain expertise and good client communication skills.

When we tried setting up a similar practice in India, we faced a challenge. The specific combination of skills required for the analytics group was difficult to find locally. It seemed like an impossible task. However, I decided to approach the situation differently. Rather than focusing on finding one individual with all the necessary skills, I broke it down into several distinct skill sets that could collaborate to achieve the desired outcome. Then we began scouting for individuals who would fit our specific requirements.

Case Study - Solving for unavailability of specialized data talent issue through Data Pods

The Genesis of Data Pods - Bringing three skills together



The Model that Scaled-up

McKinsey built-up an entire service line on this model

Although initially, we had begun by hiring statistics graduates from one of the top statistics schools in India as data scientists, we soon discovered that they alone were not the complete solution. Firstly, there was a scarcity of talent available, and secondly, many lacked fundamental problem-solving and communication skills. It quickly became evident that relying solely on statistics graduates would not suffice. So, to address this, we sought out individuals with problem-solving and communication skills, leading us to hire MBAs. At that time, several IITs (Indian Institutes of Technology, the premier engineering schools in India) had established their business schools with a prerequisite of having an engineering degree, making it feasible to recruit MBAs who possessed both client engagement abilities and a maths foundation. We also realized the importance of tasks involving data cleansing and basic data operations, which prompted us to bring in MCAs (Master of Computer Application).

This team-based approach not only solved the talent problem but created an opportunity to scale up analytics in McKinsey, which could never have happened with the traditional individual-skills focused model. Within a few years, we grew from a small team to a large, scalable analytics unit, comprising hundreds of professionals. This transformed the concept of analytics at McKinsey from a niche offering to a mainstream service line. Today, analytics contributes significantly to McKinsey's revenue.

This experience taught me the importance of being creative and adaptable when seeking talent. The creativity and problem-solving required to design a scalable and winning talent model is no less than that required for product or business model innovation. Different talent markets offer different skill sets. Unlocking the true value of talent requires going not with a one-size-fits-all approach but understanding the local talent ecosystem deeply and envisioning scalable models embracing the local talent nuances.

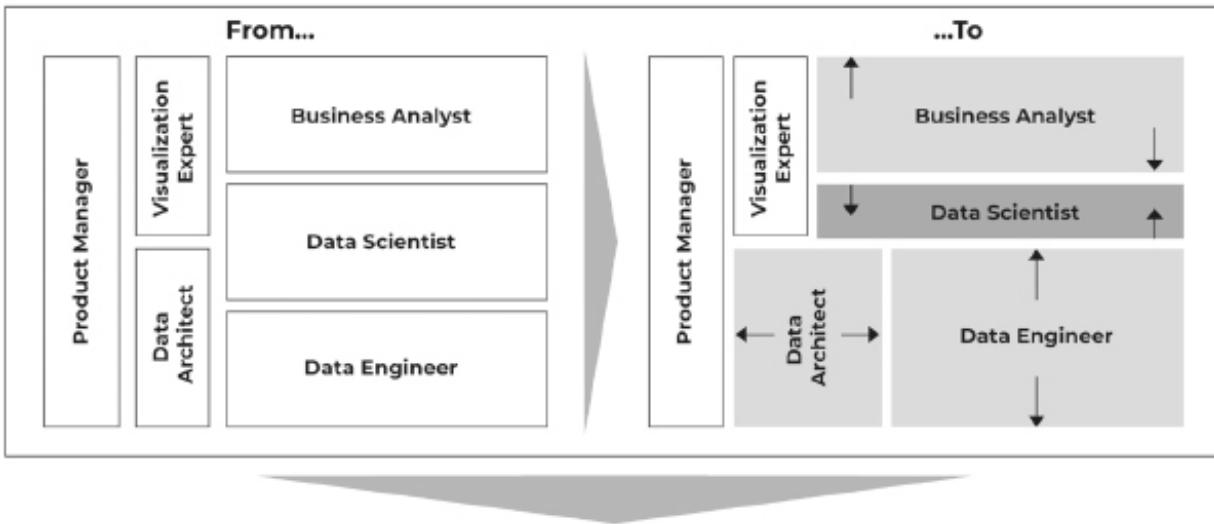
Data talent landscape is evolving in the AI age

The demands of the rapidly changing business environment and the advancement of the AI age is already beginning to bring in significant disruption to the data talent landscape as well. And therefore, I expect a significant overhaul where a few roles become more prominent, and many undergo significant change and become more specialized or narrower.

As we discussed, value is generated at each stage of the data management value chain. But in the AI age, the two ends of the value chain, the 'data' and customer/business 'impact', play a critical role in shaping the talent landscape. Data, which is growing in complexity and scale and fuels the massive AI models, holds utmost importance in terms of quality, relevance and proper utilization. So, the role of data professionals in handling such complex data landscapes is naturally expanding. On the other hand, as the importance of leveraging data effectively in driving business outcomes has increased tremendously, it has become increasingly

important for data professionals to expand their capabilities beyond core data skills to become more business outcome focused.

Data roles are going to evolve



- ① Data Science roles may shrink and become more specialized
- ② Business domain expertise, problem-solving and storytelling will gain prominence
- ③ Data Engineer and Data Architect would gain prominence

I foresee three key trends that will shape the data talent landscape:

1. Traditional data science roles are shrinking, becoming more specialized

Once dubbed the 'sexiest job of the twenty-first century', data science has now become an integral part of data-driven organizations,² and will continue to be in high demand especially with the rising importance of Gen AI. Data scientist roles are expected to increase by 35 per cent from 2022–23, making it one of the fastest-growing professions in the United States.³ While I acknowledge that this trend is likely to continue, the conventional skill set associated with the data scientist role may become less relevant, paving the way for a fresh definition of what it means to be a data scientist. Here is how things are rapidly evolving:

- **Automation and AI advancements:** The progress of AI and automation is causing many of data scientist's traditional data preparation tasks to shrink. Tasks like data cleaning, data processing and data preparation and feature creation can now be automated through tools like AutoML (automated machine learning). Approximately 80 per cent or more of a data scientist's job involves preparing data for analysis and this can be significantly automated now.⁴ The emergence of low-code and no-code platforms, which are easy to use, has simplified the data preparation process, making it more accessible and widely adopted, freeing up a data scientist's time to focus more on generating relevant insights.
- **Pre-built models on cloud:** Hyperscalers like AWS, Azure and GCP boast rich libraries of pre-built analytics models that are easily deployable with minimal customization and which streamline the process of generating insights. This significantly cuts down on the time-consuming model development process, which has traditionally been the 'bread and butter' of data scientists. As a result, these traditionally specialized tasks can now be done with minimal expertise.

The above trends are leading to the following changes in the role of data scientists:

- **The rise of 'citizen data scientists':** Traditionally, data scientists were primarily responsible for statistical modelling and building various data models. The democratization of data and AL/ML automation tools has led to the rise of 'citizen data scientists'—individuals with limited formal training in advanced analytics, statistics or related disciplines who perform data-related tasks. The availability of various pre-packaged and user-friendly data analysis and

visualization tools has made it easier for non-technical users to work with data. As a result, this role is now often played by business analysts but also many other profiles. They are expected to be more data competent and to handle tasks ranging from defining data requirements to building simple data models, which were previously the domain of data scientists.

- **High focus on specialized roles:** The need for a generalist data scientist is shrinking and their role is becoming more nuanced or specialized. In the AI age and Gen AI in particular, while consumption of models has become easier, the creation of these models is way more complex and requires sophisticated skills. These models require dealing with advanced algorithms like LLM, artificial neural networks (ANN) and NLP, among others.

2. **Business-related aspects are gaining prominence**

I believe that with the growing involvement of data professionals in driving business outcomes, they require capabilities beyond their core data skills. First, they must possess deeper **domain knowledge** related to the industry or functional domain they are aligned to, which would enable them to effectively connect data insights to real-world business problems and opportunities. Without a solid understanding of the specific business domain, data professionals may struggle to connect with the business problems and comprehend their challenges effectively. It would impact their ability to ask the right questions. Second, they must possess strong **problem-solving skills**, essential to identify and address complex data-related challenges. This is because in addition to analysing data, they are also expected to propose solutions to business problems, optimizing processes and uncovering opportunities for growth.

Another critical skill that has become paramount for a data professional is the art of **storytelling**. Data professionals are often tasked with presenting their findings to non-technical stakeholders; this has given rise to the importance of data storytellers or data translators. Strong storytelling and communication skills allow data professionals to convey complex data insights in a compelling and understandable manner. Organizations like McKinsey recognized the importance of data storytellers or data translators early on and took proactive steps to address the need. McKinsey established an academy internally in 2017 to train 1000 individuals specifically for these roles.⁵

3. **Rising importance of data engineers and data architects**

I also believe that the role of data engineers will expand substantially, primarily driven by the exponential growth of data and the engineering complexity associated with it. Add to that, the rapid advancement of AI technologies. This demand for data engineers is poised to cut across industries as organizations increasingly recognize the importance of leveraging data for decision-making. Likewise, the role of data architects will gain even greater significance in the AI age. They will be instrumental in designing and implementing the complex and scalable data infrastructure required to support AI, and now Gen AI as well. The expertise of data architects in areas such as data modelling, integration, governance and system design becomes vital for managing the complexities inherent in the ever-expanding data ecosystem.

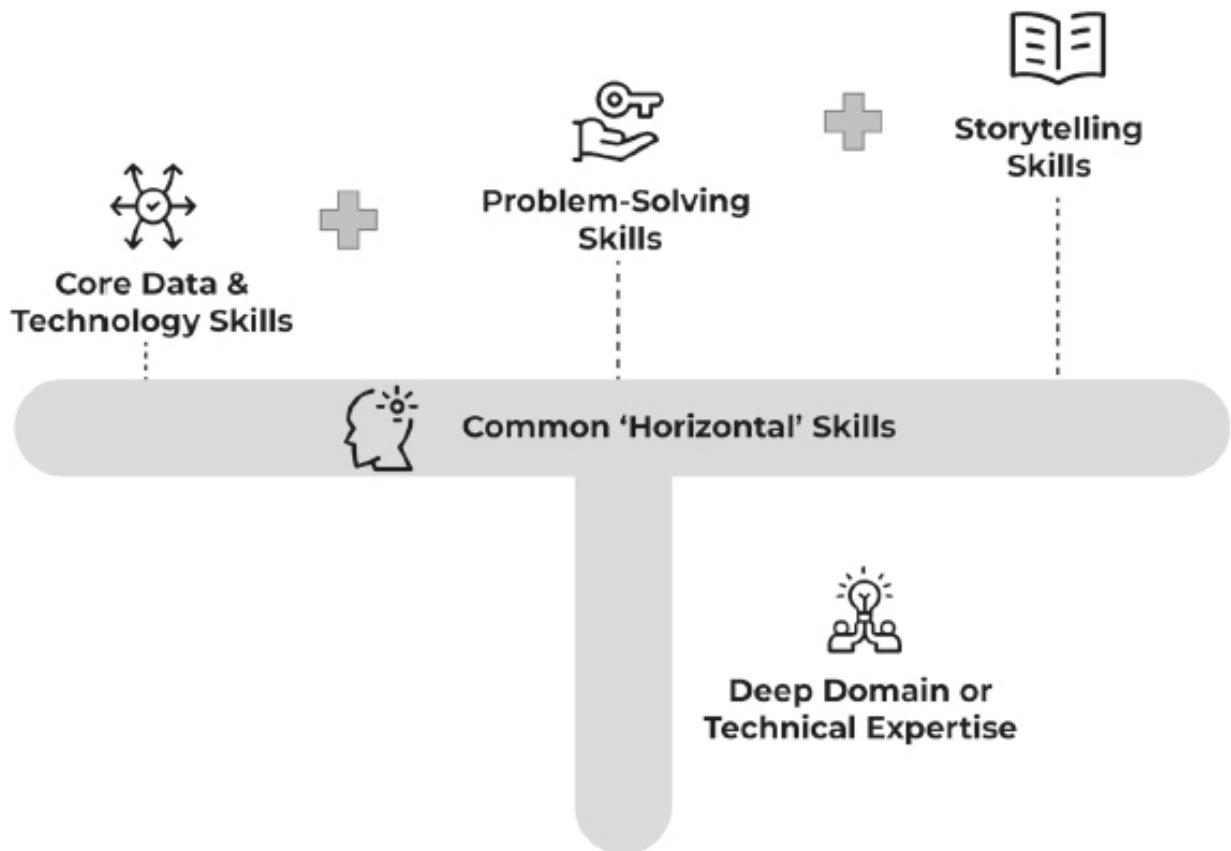
The role of data engineers and architects is gaining prominence due to four key drivers. Firstly, the **evolving nature of data** marked by its expansion across the 3V

dimensions—volume, velocity and variety—requires building robust data pipelines, storage solutions and integration processes, for effective utilization of the growing data. Secondly, with the adoption of **AI and Gen AI**, ensuring data quality becomes critical, taken care of through robust data validation, cleansing and transformation processes. They also play a pivotal role in facilitating seamless integration and optimization of AI technologies within organizations. Thirdly, data architects contribute by envisioning and designing comprehensive **enterprise data management frameworks** and act as strategic partners in effectively leveraging **new and innovative technologies** to enhance the maturity of both technical and data infrastructures within organizations. And finally, adoption of **cloud technologies** relies heavily on data engineers and data architects who are instrumental in optimizing data workflows, designing and implementing data integration processes, including extract, transform, load (ETL) workflows, for efficient data processing and storage in cloud platforms.

T-shaped capability building is the key

While there are many specialized roles making up the data talent landscape, I believe a common skills framework required by any data professional to thrive in the data-first world and the AI age is the T-shaped capability model. Every data professional must cultivate a set of essential ‘horizontal’ skills that starts with core data and technology expertise. In addition, these professionals should prioritize the development of problem-solving abilities and the art of effective storytelling. By combining strong problem-solving capabilities with the ability to tell compelling stories, professionals can enhance their contribution in driving business outcomes.

T-shaped capability building is the key for data talent



Along with these foundational horizontal skills, data professionals must also build a certain depth, be it deep domain knowledge or deep technical expertise. Deep domain knowledge involves becoming subject-matter experts within a specific industry or function, allowing them to understand the unique challenges, trends and nuances that data can address. On the other hand, deep technical expertise involves building deep understanding of technology solutions, mastering advanced data engineering techniques, and more, enabling professionals to build strong capabilities, develop advanced solutions and work with complex data sets effectively.

This combination of horizontal data skills and specialized depth equips data professionals to not only be effective with data but also to provide tailored insights and solutions that drive meaningful

impact within their respective domains, whether it's healthcare, finance, marketing or any other sector.

Key takeaways

- Specialized data talent is crucial for organizations to extract transformational value from data. Data talent consists of multiple roles like the business analysts, data scientists, data architects, data visualization experts, data engineers and product managers. These data roles are significantly evolving in the AI age.
- The traditional role of a data scientist is shrinking with the advent of AI and automation, availability of pre-built solutions and expanding role of business analysts and emergence of 'Citizen Data Scientists'. This has led to the data scientist's role becoming more specialized.
- Data engineers and data architects are becoming even more critical in managing the data complexities inherent in the AI era, owing to the evolving nature of data, increasing AI and Gen AI adoption, innovation in the technology landscape and cloud adoption.
- To thrive in the AI age, every data professional must develop T-shaped capabilities. It is necessary for every data professional to have core data and technology, problem-solving and storytelling skills. These should be supplemented by building depth either through deep domain knowledge or deep technical expertise.



SECTION III

DATA FOR INDIVIDUALS AND BEYOND

'As is the microcosm so is the macrocosm, as is the macrocosm so is the microcosm.'

—Yajurveda, ancient Indian scripture

Moving from Enterprises to Individuals and Society

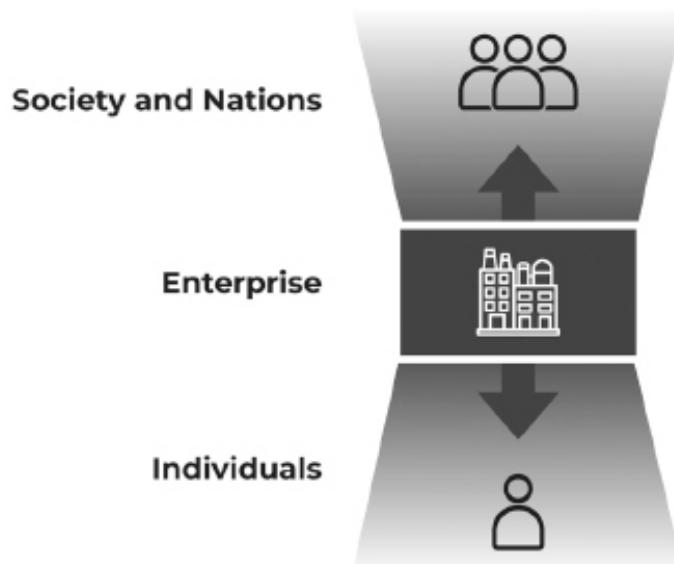
'As is the microcosm so is the macrocosm, as is the macrocosm so is the microcosm.'

—Yajurveda,
Ancient Indian scripture

I am a firm believer that at a fundamental principal level, what holds true for a microcosm, is true for macrocosm as well. This concept has its roots in the Indian scripture called Yajurveda, where in Sanskrit it is said: '***Yatha pinde tatha brahmande, yatha brahmande tatha pinde***', which means that as is the structure of the microcosm, so is the structure of the macrocosm and as is the structure of the macrocosm so is the structure of the microcosm. This is the core philosophy of internal inquiry in the Yajurveda which states that the entire universe, in fact everything that exists outside our bodies, exists inside our bodies too. This essentially means that the fundamentals across various levels of existence are the same.

This powerful concept holds true for data as well. I believe that the data concepts that I talked about in Sections I and II in the context of an enterprise, are equally applicable for individuals, the society and nations. While the execution or scale may vary, the cause and effect remain similar. And if we start looking closely, we will find patterns and similarities which prove that what holds true for enterprises, also holds true for individuals and vice versa. These patterns and similarities are the essence of a lot of my work.

Transitioning from enterprise to individuals and beyond



So, this section is dedicated to demonstrating how the previously discussed concepts, applied in the context of enterprises, are equally relevant and valuable for individuals, society and nations.

Some of the challenges that I have laid out in this section for individuals, society and nations, may not have a simple and ready solution. I do not claim to have the answers on how to solve each of these complexities, but I do have a few mantras which come from my experience, that can empower both individuals and nations to make more informed choices.

And to begin with, 'As is the macrocosm so is the microcosm, as is the macrocosm so is the microcosm' is one of the first mantras that I strongly believe will help you make sense of all the data chaos around you by helping you identify the patterns well. The concepts, once understood at an enterprise level, will find application at an individual level and vice versa.

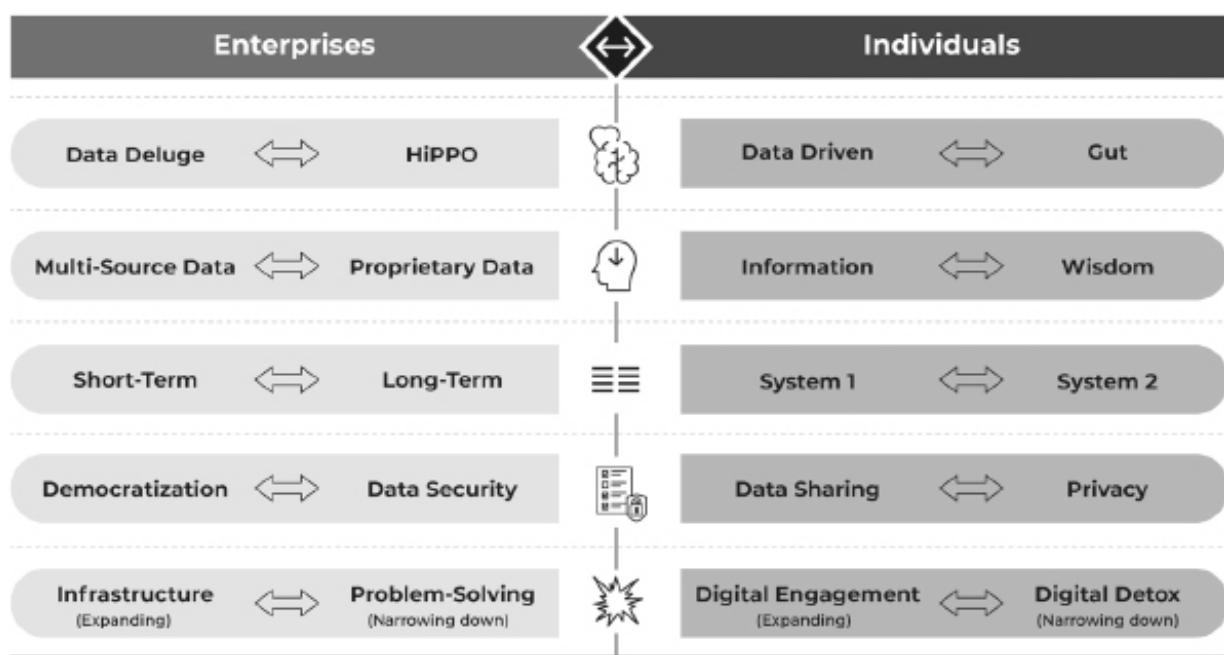
If we understand this philosophy, we will start finding opportunities and solutions. We will start looking for patterns. For example, executives who are trained to use data in decision-making in their workplace are often seen relying on their gut, when it comes

to personal decisions. Historically, this was because there was no or limited data available at a personal level. But as the world entered the age of smartphones, access to data became easy and cheap for individuals. Yet, in the data-abundant world, there is very limited usage of data in personal decision-making. We will see how we can apply the learning from the concepts I talked about at enterprise level, for better decision-making at the individual level as well.

The Data Paradox also manifests itself for individuals

In Section I, I highlighted that the phenomenal growth of data has brought about unprecedented opportunity for organizations to derive transformational value from data. At an individual level as well, this abundance of data comes with a promise of delivering hyper-personalized experiences. Today, products and services can be highly tailored to meet the distinct needs and expectations of each person. Furthermore, data that has completely transformed the way we live today, brings with it the tremendous opportunity to use it to enhance our well-being through data-driven decision-making.

Enterprise-level paradoxes are applicable at an individual-level as well



Having said that, just like an organization, individuals also confront several paradoxes due to this data abundance, hindering their ability to fully harness its transformative potential. While the essence of these paradoxes is similar, they often manifest in different forms for individuals.

The overarching concept, the Data Paradox, deluge vs drought, that inflicts an organization, also manifests in individual decision-making. In a world where information is pouring in from all directions—whether it's from the television, smartphone, internet or wearables—individuals find themselves constantly bombarded by a relentless deluge of data, every second of every day.

The digital world never sleeps. Most times, individuals are overwhelmed by it all. On the other hand, how much of the information really does go into decision-making at personal levels? Most of us still rely on our gut and experience, or the experience of our trusted ones, to make decisions in our lives, small or big! Just like turning to data helps organizations diminish some of the ill-effects of HiPPO-based decision-making, individuals must also find ways to effectively incorporate data in their decision-making process. Moving away from always relying on the gut and using data helps make informed decisions. Biases are removed. One does live better when one includes data in their decision-making process.

While data-abundance may seem like it empowers individuals in decision-making, the paradox is whether this abundance truly enhances wisdom or merely creates a facade of intelligence. The interplay between information and wisdom echoes the enterprise dynamics of multi-source data and proprietary knowledge. Wisdom, like proprietary knowledge, sets individuals apart and enables them to make consistently wise decisions. So, is data abundance really making us smart or is it just creating a false sense of being smart? Does data abundance create a paradox of wisdom? Does it lead to more wisdom or less? The manifestation of this Data Paradox is even more stark at an individual level than in an enterprise.

Individuals, similar to organizations, have to survive and thrive in the highly dynamic, fast-changing world. Hence, just like the two-speed approach that organizations have to adopt to be able to

achieve short-term goals while building long-term capabilities, individuals have to connect System 1 and System 2 thinking (System 1 thinking is a near-instantaneous process; it happens automatically, intuitively, and with little effort. System 2 thinking is slower and requires more effort). They must deliberately adopt System 2 thinking, which is based on logical reasoning and analytical thinking to make informed decisions. In the long run, when exposed to repetitive scenarios, the process becomes intuitive, strengthening our System 1 instincts!

Another central question that rises out of the data revolution is, do you participate in this revolution or not? The answer is simple—you ARE participating, knowingly or unknowingly. Every time you use your phone or access the internet, you leave a digital footprint. It's no longer a choice. Therefore, the key challenge lies in finding the right balance between data sharing for personalization and safeguarding one's privacy, similar to the balancing act that an organization has to perform between democratization and security.

The ubiquitous nature of digital engagement introduces another paradox similar to the expanding infrastructure vs narrowing down to solving a problem effectively. As organizations invest their time and resources in building expansive infrastructure in the hopes to solve their business problems but get caught up in that process, individuals also start accumulating digital experiences in an attempt to enhance their lives and live better. While such digital engagement does simplify daily tasks and complement modern lifestyles, it often introduces challenges such as feelings of inadequacy, depression, anxiety, digital echo chambers, cyberbullying and much more. Narrowing down to clearly identify personal goals and objectives, and focusing on specific, well-defined objectives or challenges, requires regularly stepping away and connecting with one's inner self to establish a fine balance.

These paradoxes can have distinct effects on organizations versus individuals. For organizations, they may result in anything from generating inaccurate insights to completely immobilizing the entire operation, bringing it to a standstill. For individuals, the impact of

paradoxes can vary from erroneous decision-making to severe mental health problems.

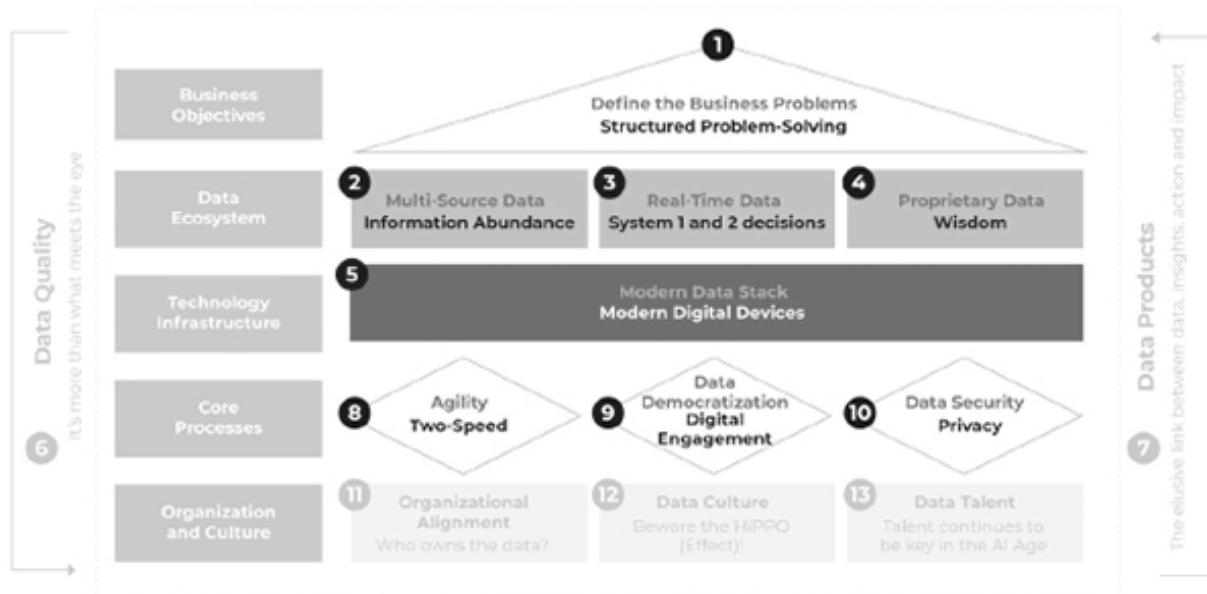
Furthermore, in this section, I have also highlighted parallels when I look at the nations of the world and the society at large. Can data collaboration help solve some of the greatest problems we face as humanity, for example—healthcare, climate change and lead to a better, more sustainable world? Can data support in delivering prosperity? Can data help nations in achieving competitive advantage? Can data support in delivering prosperity? If so, how? What does a government have to do? How is what a government does different from what the enterprise does? What can governments learn from enterprises to compete on a world stage? What is the role of citizens in this?

By the way, global collaboration and national competitive advantage can be seen as contradictory to each other. However, both are important objectives and have to be solved simultaneously. This is another paradox where you can't just chase one objective. Making sense of this and more such paradoxes require recognizing and learning from the patterns between the microcosm and macrocosm!

The USF can become your personal guide

Interestingly, when I look at the thirteen-component Unified Solution Framework talked about in Section II, I do see a parallel application of many of its key components in the individual context as well.

The Unified Solution Framework (USF) for Individuals



Define the business problems: A structured problem-solving approach holds true, whether we are solving a problem for an individual, enterprise, society or government. A systematic approach is required to clearly define the problem, break it down into components in order to narrow down the core issues to be solved. It makes the entire process more manageable and impactful.

Multi-source data: Just like enterprises, individuals today have access to data from multiple sources, resulting in **information abundance**. Thus, for individuals as well, the ability to triangulate the information from multiple data sources becomes a key enabler in better decision-making. For example, while deciding on the medical treatment, typically we would like to take the opinion from multiple doctors to triangulate and arrive at the right course of treatment.

Real-time data: Just like real-time data refers to information that is collected and processed immediately as events occur, **System 1** is a fast and automatic type of decision-making that humans engage in, often based on intuition, heuristics, or instinct. While **System 2** involves careful analysis, rational thought and conscious decision-making. The choice of which system to use depends on factors such

as criticality and time constraints. While some decisions may be ideal for System 1, like braking at a red light, in some cases one has to engage in more deliberate analysis (System 2) to make a final decision, like deciding which car to buy!

Proprietary knowledge: For organizations, codification of tacit knowledge is the key to build their proprietary knowledge. Similarly, for individuals, it is reflecting, observing patterns and connecting the dots to build wisdom in the long run—which I believe is proprietary to an individual. Just like proprietary knowledge, wisdom is not easy to codify and thus does not extend beyond the individual, but equally it is very important that it is codified and shared.

Modern data stack: Just as a modern data stack is crucial for translating data into insights and making it available for decision-making, engaging with **modern digital devices** empowers individuals to transform data into meaningful insights. Both scenarios involve leveraging new technology to harness the potential of data for informed decision-making and understanding the world around us. These modern digital devices serve as powerful tools for individuals to collect, analyse and derive insights from data.

Agility: The **two-speed** approach, as critical for organizations to survive and thrive in the Big Data world, stands true for individuals who are living in a fast-changing, dynamically evolving world we are in today. So, making small changes that can be implemented quickly and easily to get immediate results, while recognizing and working towards those that can be sustained over the long term to become habits, is a key to living a full life.

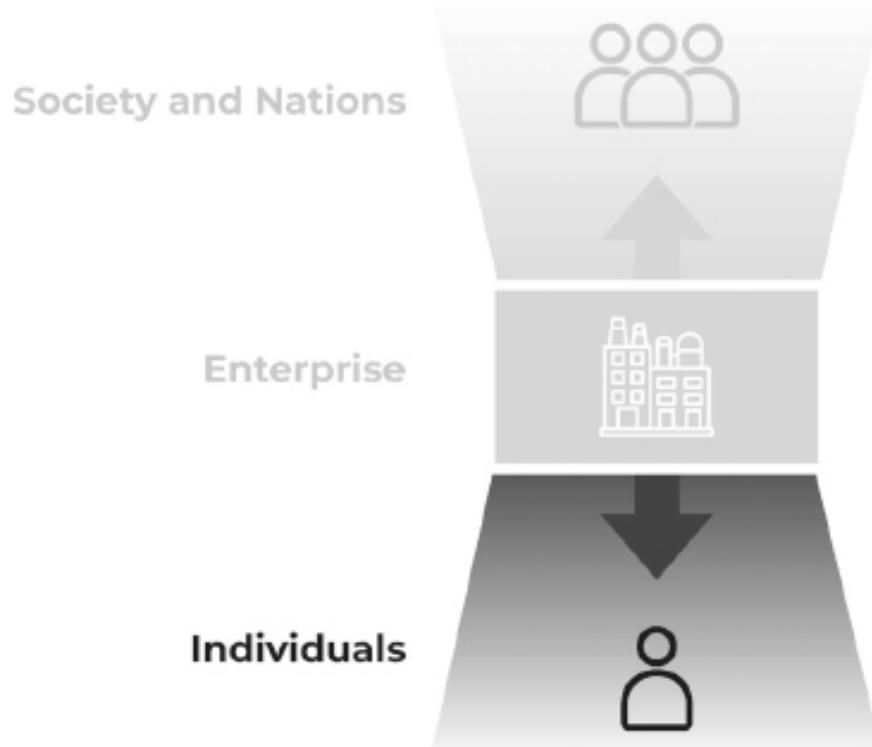
Data democratization: While organizations need to make sure that they democratize data access for their teams and build capabilities, individuals are waking up each day with access to data, democratized by the information available on the internet and increasing penetration of wearables and multiple other sources. This has led to a significant shift towards **digital engagement**,

predominantly moving away from offline mode. These devices democratize access to data and analytical capabilities, empowering individuals to make informed decisions, improve their personal lives and contribute to various fields and industries. Gen AI is already beginning to play a significant role by making the data of the world available to all, to create interesting insights for individuals, enterprises, society and governments, if contextualized well.

Other concepts, like data security seeing parallel in data privacy, and data culture translating into data mindset, reinforce my belief that the USF framework is not just for enterprises but has applications across.

The next few chapters are my attempt to elaborate on the opportunities of the Big Data world for an individual, the society and the nation, address the most critical of the paradoxes that individuals face and answer a few burning questions and more. I hope to provide you with a few practical tips to harness the power of data, to live better as an individual, collaborate as a society and build a competitive advantage as a nation.

INDIVIDUALS



'Data transforming individual lives'

The World of Hyper-Personalization

Segment of One

'Personalization is when you walk into a bar and the bartender puts a whiskey in front of you, without having to ask what you want.'

—Jeff Bezos,
Founder and former CEO, Amazon

As people who have gotten used to the internet, haven't we all become a little spoiled? Eager to find answers to all our questions at the click of a button? We turn to the world wide web to get an immediate answer to a question that we are looking for and Dr Google to help diagnose our symptoms with the click of a few keys. We now expect entertainment to be on-demand, binge-watching our shows vs waiting for days for the next episode to release. We want Netflix to make shows we like, Amazon to give us recommendations on what to buy, Instagram feeds to be filled with recommendations on the kind of content we are watching. Each one of us has, to some extent, been spoiled by the personalization being offered by the digital natives or disruptors as we call them—what we want, where we are served, at times even without asking. And we want traditional organizations, big or small, to follow suit. This is the new data-first world we live in today.

But this wasn't the case until just a few years ago when we were catered to as part of a segment and not as an individual. We were all put into different groups, who experienced similar products and services and had to find our own individual ways to personalize it. While this approach succeeded to some extent, it didn't turn out to

be the most effective way to serve customer needs or influence their choices. Let me explain why.

The fundamental flaw of traditional segmentation

The need to know the consumer intimately was understood even in the 1980s, but there were limitations. Organizations had no way to collect such extensive data on each and every consumer to be able to successfully build products/services tailored just for them. So, the next best option was the logic used by marketers, that people could be grouped into segments based on similarities in their demographics, location, choices, etc.

Although it did succeed to a certain extent in narrowing down the target segment, I believe this approach was fundamentally flawed. Why? Because it did not take into account an inherent human trait—each one of us is unique in our own ways and wants to be treated that way too. Ironically, in the billions of people around the world no two human beings have the same fingerprint or DNA, yet for years, we were treated as groups—segments!

But this is not so much by design but by necessity. The organization of the 1980s and 1990s had limitations of data and technology. Hence, they had no better way to bring in some level of targeting of customers through segmentation. At that time, this was the best way to create solutions, to give consumers what they needed.

A glimpse of *Mad Men*, a popular TV show, is enough for anyone to understand the inner workings of the advertising industry in the 1950s, portraying the power of head honchos in controlling the consumer narrative and the choices. Advertising campaigns were designed to evoke strong emotional connections, making consumers feel part of a larger narrative associated with a brand, which helps influence the customer purchase behaviour. The show portrays the era before extensive use of data for market research became prevalent, when the advertising executives were more focused on influencing the customer behaviour, rather than understanding it. Yet, there were thought leaders like Howard Sharman—editor of

Marketing Week from 1984–88, and publisher from 1988–93—who observed that ‘through the 1980s, the brand owners slowly gave up the pretence that they owned the consumers that bought their products’, eventually accepting that an advertising campaign was not a guarantee to raise sales. So, they moved to below-the-line advertising—promotional activities, like discounts, etc.

But all this was not so much by design but by necessity. The organization of the 1980s and 1990s had limitations of data and technology and there was no way to know what a consumer exactly wanted. So, during this industrial revolution era, personalization remained a domain of the rich. It was a luxury. Organizations, even though they desired to offer personalized customer products, were not geared to offer the same, as they did not have a way to know what each individual wanted.

However, the thought process of trying to know the consumer intimately, started in the 1980s, is seeing its fructification now, in this era of data abundance, when the power is shifting back to the hands of the consumers. As the data revolution started making inroads, individuals started sharing data and the world moved from offline to digital, organizations now had the ability to create offerings which were relevant to each person, individually. It was a journey which took many decades of evolution in the business world.

Emergence of segment of one: An outcome of data abundance

When each person is addressed as an individual and not a part of a group, it is a ‘segment of one’. David Edelman of BCG defines segment-of-one marketing as a method that brings together two formerly independent concepts: information retrieval and service delivery. On one side is a proprietary database of customers’ preferences and purchasing behaviour; on the other, is a disciplined, tightly engineered approach to service delivery that uses the information base to tailor a service package for individual customers.¹

The emergence of the ‘segment of one’ represents a significant shift in the realm of personalization and customer experiences. The concept refers to the idea of tailoring products, services and content to suit the unique preferences and needs of each individual customer, contradictory to the traditional mass-customization or segmentation strategies.

This transformation has become possible primarily due to the explosive growth of data across various dimensions, as discussed in Section I. Alongside the increase in the volume of data available for individuals, organizations have also seen a significant expansion in the granularity of data they can collect. This means that there is now a wealth of diverse data points that can be harnessed to gain a deeper understanding of each individual. Consequently, organizations have the ability to accurately predict the needs of each individual and tailor their offerings accordingly. Furthermore, real-time data collection on individuals aids in comprehending their current behaviour patterns and recommends solutions based on these observations in real time.

My meeting with the head of data science of Verizon in 2018, led to a revealing comment to my simple question: ‘How do you define your segments?’ Without batting an eyelid, he said, ‘I have 140 million segments.’ I was initially stumped by his answer, but soon realized that he was talking about each one of Verizon’s 140 million customers. It was then, in that moment, that I realized the power of the ‘segment of one’.

At an individual level, in the telecom sector, this translates into personalized offerings where you can choose any plan you need, based on your specific usage and reduce waste. You can manage your budget, and track and monitor your usage. The control of how much data you use and what you pay for it, is in your hands, not the company’s. This shift in control of decision-making from an organization to an individual is possible and is happening due to data abundance. Organizations now have the ability to service the segment of one. Individuals now have the luxury to expect a service, which is for them only. According to a survey, 71 per cent of consumers anticipate personalized interactions from companies, and

76 per cent express frustration when this expectation is not met. And companies that excel at doing so, are known to generate 40 per cent more revenue from personalization, compared to their slower-growing counterparts.²

In a survey of 300 wealth-management executives, 82 of respondents believed that only those who increase product personalization will succeed.³ Today, almost every investment adviser's website has an online SIP calculator. This is a classic case of a personalized offering, as within minutes of a prospect entering details about your requirements, you are presented with a customized SIP plan.

Hyper-personalization is swiftly becoming a fundamental aspect of how organizations provide value, and consumers are progressively anticipating personalized experiences across all facets of offerings. So much so that nearly three out of four consumers say they would **only** buy from brands that offer personalization.⁴

Breaking a long-standing paradox with hyper-personalization

During the industrialization era, as society experienced unprecedented advancements in technology, manufacturing and urbanization, we moved from personalized, bespoke, individualized products and offerings, towards standardization and scale. For very long, we lived with the paradox between individuality and scale. The method to achieve scale was seen as standardization. So historically, personalization and scale were polar opposites, and we had to sacrifice one for the other. You could either ask for personalization, or had to live with standardized, common for all products and offerings. Henry Ford, owner of one of the automotive giants of that era, is known to have famously declared, 'Any customer can have a car painted any colour that he wants, so long as it is black.'; a fair representation of the thought process that we lived through the pre-Big Data era.

Now with Big Data and the dawn of the AI age, we are at a pivotal moment in human civilization. We are presented with an unprecedented opportunity of breaking that paradox. We finally have the means of achieving individuality at scale! Now with data and AI, we have the ability of providing personalized, bespoke offerings to each and every individual, while catering to millions of individuals at the same time.

That is the power of the data-first world and the AI age!

Digital, Data and AI: The enablers for segment of one

Earlier in this chapter I have already talked about how abundance of data has enabled the segment of one. Another important enabler is the emergence and widespread adoption of digital technologies that have transformed every aspect of our lives. Technologies including smartphones, high-speed internet, wearables and other IoT devices have collectively revolutionized the way we communicate, work, access information and conduct business. And the more we engage with these technologies, the more data we consume and create on a continuous basis, giving way to the explosive growth of data available today. I have mentioned this virtuous cycle between Digital, Data and AI in Chapter 2, Data, the Fuel for the Digital Age, as well.

Today, there is a constant stream of near real-time and real-time data that you are sharing with each of your clicks on the phone and keyboard, that gives enough signals to an organization to customize their engagement with you.

Have you ever wondered how, when you put an item in your check-out and the item is not bought, you get a reminder or a call from the company that you have a purchase waiting to happen. And very often, that call, or reminder happens within minutes of you abandoning the cart. This is the power of near real-time data that the companies are collecting about you, to nudge you or service you better.

AI combined with the increased power of computing helps interpret and take action on the signals that real-time data is sending. With AI, organizations can now deliver hyper-

personalization at scale, in the moment. The brands now have the power to provide customers exactly what they want and when they want it. The concept of 'Amazonification' exemplifies how organizations like Amazon have harnessed data-driven recommendation engines, AI and analytics to understand and predict customer buying behaviour. This, in turn, has transformed the online shopping experience, making it more convenient and personalized. It's similar to having an online salesperson who uses subtle cues from your interactions to suggest products or content that align with your preferences.

In the content world, we call this Netflixization, where platforms like Netflix use data and unique algorithms to recommend content, based on an individual's viewing history and preferences. This approach minimizes the need for extensive searching and instead offers tailored content suggestions.

Co-creation is key to achieving effective segment of one

'Co-creation' is key! It involves an active collaboration process between users and organizations to create highly personalized experiences, maintaining a sense of individuality and uniqueness.

Organizations of every era have tried to co-create and work with their consumers. However, in the Big Data world, organizations can do so by gathering data around preferences and behaviours, made possible by the ubiquitous presence of digital in an individual's life, generating multiple data points on each individual through their online interactions, purchase histories, social media presence and more. Organizations then use advanced analytics and ML algorithms to analyse this data, identifying patterns and trends. This allows them to create personalized products, services and experiences by tailoring recommendations, content and offerings to match each individual's unique preferences and behaviours.

Co-creation empowers individuals to have an active role in shaping the offerings they receive, fostering a deeper sense of ownership

and satisfaction. By involving customers in the design process, businesses can deliver personalized solutions that precisely match individual needs, leading to enhanced customer engagement and loyalty in the hyper-personalization era.

A good case in point is the Nike Sport Research Lab. How are they so successful in designing apparel that enhances an athlete's performance? It's through co-creation. They bring in athletes to the lab and run a number of tests on them. With that data, they are able to build deep knowledge on a large number of athletes in the world. Armed with this knowledge, they explore numerous ideas and conduct extensive testing in pursuit of improving performance, reducing injury risk, enhancing perception and feel, and delivering innovative products to athletes.⁵

Another example is LEGO Ideas, which allows Lego fans and enthusiasts to submit their own original LEGO set designs to the platform based on movies, TV shows, famous landmarks or any creative concept. Ideas that garner more than 10,000 supporters among the Lego community, stand a chance to become an official Lego product. This process not only fosters a sense of ownership and community engagement, but also enables LEGO to tap into the creativity and ideas of its passionate fan base.⁶

A personalized offering can only be created when the user and the provider collaborate, share information—data. In the past, very few companies have been successful in co-creating because it required innovative approaches to get access to relevant data. But now, owing to the proliferation of digital technologies, organizations can engage with customers on a one-to-one basis more effectively than ever before. This has led to the creation of highly relevant and personalized experiences, and it has become increasingly common across various industries.

Even within highly regulated industries like financial services, the use of data is driving efforts to provide hyper-personalized experiences. A notable illustration can be found in Erica, the virtual financial assistant offered by Bank of America. Erica engages in personalized interactions with customers, drawing insights from

various sources like account balances, transaction history, spending patterns, payment alerts and duplicate charges to tailor its recommendations and conversations to individual customer needs and preferences.⁷ Similarly, a UK-based fintech company Cleo uses open banking transaction data and deep-learning technology to provide personalized recommendations to clients through a chatbot. Users can seek answers to queries like 'What's my monthly spending on groceries?' or 'How can I save £300 by the end of this month?' Furthermore, the app can offer specific guidance for users to achieve particular life goals, such as creating a savings plan for an upcoming vacation.⁸

Knowingly or unknowingly, the individual is co-creating the solution. Sharing the data allows the organization to map the information with the enormous amount of data they already have on the individual and others in the market and carve out a potential solution that is personalized.

Having said that, co-creation is most effective when customers/individuals participate willingly in the process, providing accurate and up-to-date information. Because this data acts as a foundation for organizations to design and enhance their products, services or experiences to align as closely as possible to an individual's unique preferences, needs and wants at all times.

Active participation of the users in the personalization process offers them control and transparency and allows organizations to navigate the delicate balance between catering to individuality and achieving scale in their hyper-personalization efforts. This approach respects users' uniqueness while providing tailored experiences that align with their preferences and values.

Segment of one shifting the power of choice to the individual

Hyper-personalization has brought in a shift, when the power of making a choice has shifted back into the hands of the individual. I call this a shift back, like a pendulum swing, because it harks back to

the pre-industrial revolution, when each individual created or bought products customized for himself. From clothes to food, everything was customized. Today, thanks to the data revolution, organizations have developed a capability to offer experiences highly relevant to an individual, while still mass producing. As oxymoronic as it may sound, ‘mass customization’ is a reality today. And as data continues to explode, and individuals share more and more data, the segment of one is becoming increasingly attainable.

As is evident from the various examples of hyper-personalization, the value of hyper-personalization is typically in—improved choices, higher relevance, higher value and more engaging content. As we are getting pampered by organizations, we are adapting very quickly and gaining benefits from being that segment of one. We are now expecting these benefits from any of the interactions we have with a brand or an organization.

As people share their data, they are able to get solutions and options that are customized to them. These solutions are different from their neighbours. The realization that they are unique is higher than earlier. And the realization that they can get what they want is further driving their expectations from organizations as well.

Segment of one is where every individual’s uniqueness is fully recognized and addressed. And by treating each individual as a unique segment, organizations can optimize their offerings and recommendations to meet the exact requirements of each customer.

Benefits of hyper-personalization for individuals

Improved Choices

Provides options that are **customized based on individual's preferences**



Higher Relevance

Recommendations on content and offers with higher relevance for the individual

Higher Value

Deliver higher value as hyper-personalized solutions **target individual's specific pain points & needs**

Higher Engagement

More engaging **content, products, offers and recommendations**

Let me again use Nike as an example to showcase the benefits of using data to cater to segments of one. Nike is using an app that provides access to Nike plus rewards programme which provides access to offers, exclusives and early access to new products. It also provides access to experts that recommend products based on your requirements. By crunching the vast amount of customer data from the Fit app and other wearable devices, like Fitbits, Nike gains insights into customer habits and can predict their purchasing decisions more effectively.

Additionally, using the Nike Fit app, customers snap a picture of their feet using their phone and get the perfect shoe size for every style of Nike shoes. The app utilizes a combination of computer vision, data science, ML, AI and recommendation algorithms to accurately measure the complete shape of both feet, enabling customers to discover their ideal and perfectly fitting shoe size.⁹

Furthermore, although made available only to VIP customers, like athletes, runners and celebrities, Nike By You studio, uses a voice-assisted environment and state-of-the-art technologies to enable customers to custom design their own shoes, within sixty minutes, which are then delivered to them in six–eight weeks.¹⁰ For the regular Joe, NikeID, an online offering, enables customers to select

various colours and materials for their shoes, and personalize them further by adding their name and/or number through stitching.^{[11](#)}

They have also been experimenting with cutting-edge technologies like 3D printing for a while, to customize their offerings for athletes and celebrities and other top players. For instance, in 2017 they partnered with Eliud Kipchoge, one of the fastest marathon runners in the world, to reduce the weight of his running shoes by 3D printing a plastic upper instead of fabric and customizing the shape to fit him better. With his lighter, more flexible and highly customized shoe, Eliud Kipchoge went on to win the London 2018 marathon.^{[12](#)}

More and more organizations are realizing the value of personalized experiences. To achieve the segment of one, most organizations have identified their ability to capture and leverage customer data as the key to drive personalization and hence are investing on building ‘customer data platforms’—a market that is poised to grow at a CAGR of 32.4 per cent to reach ~USD 20 billion by 2027.^{[13](#)} Hence, more and more businesses are investing in capturing and utilizing customer exhaust data—the substantial amount of data that is generated as a byproduct of various digital interactions and online activities, while each individual is supporting this activity, by offering their data, feedback in various interactions—online and offline.

The future of segment of one

As I look at the shifts in the world happening due to data, I can foresee a significant impact of ‘segment of one’ on every industry. However, I feel particularly excited about the prospects in healthcare and education industries that hold substantial promise for leveraging the power of data to address and cater to the uniqueness of every individual. These two industries have the potential to leverage data to cater to the uniqueness of each individual and have a very fundamental impact on human lives. This is because healthcare treatments involve dealing with a physical body, where each body is

unique, and each individual's DNA distinct. Similarly, each mind is unique, and every individual comes with unique potential. The education system will continue to play a crucial role in shaping these unique minds. Hence, I find the opportunity that data created in these two spheres of human influence, most fascinating.

With respect to education, today, each person can choose a vocation they are meant (or born) to pursue. The era of mass production is coming to an end—where schools were churning out people who were clones of each other, in terms of knowledge and skill sets. People no longer need to be 'another brick in the wall', as passionately emphasized by the famous song, 'We don't need no education, we don't need no thought control', by Pink Floyd, which was a reflection on the time, when the purpose of education was to create people who were apt to work in the industrializing world. Sir Ken Robinson, a renowned educationalist, did a fascinating piece on how the education system is broken in his famous TED Talk.¹⁴ He says, the education system was designed to make students into professionals for the industrial era, stamping similar people together and killing their creativity and individuality, but despite knowing the perils, we did not have a solution. But now, as we move into the AI age, and automation becomes more prevalent, what would set us apart is that same uniqueness—our creativity and our individuality.

The solution lies in an ancient Indian practice of *gurukul* or apprentice system. In ancient India, the learner used to stay in a residential school—called *gurukul*—each assigned a 'guru' who nurtured their skills, values and an approach to life. The education was personalized with a mentor around you. Likewise, in the west, the guild system offered structured apprenticeship programmes, where aspiring individuals learned their trades under experienced masters, progressed to the journeyman stage for independent work and ultimately aimed to become master craftsmen with the privilege of running their workshops.¹⁵

I do wonder, with data abundance, can this personalized learning and apprenticeship model be replicated, even to a limited extent?

With data, we have the ability to personalize and even be predictive about what a person should learn, depending on capabilities and aspirations. Today, each person can express their individuality in the work they do and how they choose to earn their livelihood.

The world is shifting, becoming more dynamic. Earlier, people used to stay in an occupation for thirty years, doing the same jobs. Today, skill upgradation is needed every three to five years. New industries are emerging and the past ones are getting redundant. How to know what is best for you? How to know what you will do well in? How do you know what will have longevity?

You may be an engineer but if you want to become a chef, today, it is possible. You can learn from the best chef in the world via online classes. Today, there are universities, which are 100 per cent online and helping people make career choices instantly. One is able to personalize the career they choose to be in, via self-learning. Organizations have started valuing skills over degrees in tech hiring, and hackathons are becoming a popular choice to test people for skills. After decades, people now have the means available to become who they want to be, and the only thing stopping them is their own mind. This is the power of hyper-personalization in education brought about by abundance of data, and it is happening now.

Similarly, in healthcare, advances in the Human Genome Project are enabling people to make medical decisions that are aligned to their bodies. They now have the option to go beyond generic treatments, and hope and pray that it works, to opt for a more customized course of treatment tailored specifically to their genetic makeup.

In Chapter 2, Data, the Fuel for the Digital Age, I also shared how personalized medicine is growing at a high single-digit rate annually, and the role of genome sequencing in achieving this. The Human Genome Project that unlocked extensive data points (~19,000–20,000 genome for each individual) has made personalized treatments possible. [16](#)

One of the applications has been in the field of oncology, which is at the forefront of using genomics to help develop more effective personalized cancer treatments. Genetics tests are being used for predictive analysis and to ascertain susceptibility on being prone to a disease, enabling pre-emptive treatments to begin. Now doctors can take a sample from the tumour of the patient, and after comprehensive genomic testing, running data analytics, they can identify abnormalities in the tumour, mapping this data to multiple people, who have undergone treatments, personalized and targeted treatments can be delivered. The technology and data have the ability to offer personalized treatments.

The technology is there, but it needs to be supported with people sharing their information—their private genetic code. One needs to be able to co-create a treatment protocol with the medical practitioner, based on the data shared by the patient. If the patient chooses not to share their genetic information, they cannot benefit from the advancement in medicine and the benefits that come with it.

Will people share that information to get a cure? Market pundits say 'yes', as the mRNA cancer vaccines and therapeutics market is growing at 13.3 per cent CAGR.¹⁷ The choice is with you, as an individual, if you would like to support such advances in technology to better your life.

All this is possible due to advancement of our understanding and our ability to generate data, collect it, share it and analyse it, and then take actions that are beneficial to extending human life.

Key takeaways

- Traditional segmentation, while practical in its time, was fundamentally limited as it overlooked the uniqueness of individuals. It is now possible to factor in the uniqueness of each individual because of the abundant availability of data in the Big Data world.

- The rise of the 'segment of one' signifies a pivotal shift from the long-standing push towards standardization and scale, bringing the power of choice back to an individual.
- The interplay of Digital, Data and AI are the key enablers that are helping organizations make choices on behalf of individuals by identifying patterns and preferences based on shared data, resulting in highly personalized and efficient experiences for individuals.
- Co-creation offers an effective solution to the challenge of achieving individuality at scale in hyper-personalization. It is a collaborative process where users and organizations work together to craft personalized experiences, preserving individuality and uniqueness.

Data for Better Decision-Making

Live Better with Data

'Without data, you're just another person with another opinion.'

*—Edward Owning,
Engineer, statistician, professor, author and consultant*

In a survey conducted by faculty members of the Ohio State University, US, respondents revealed that more than half of Americans primarily use gut feeling for their decision-making.¹ It is not unexpected yet revealing that despite being surrounded by so much data, very little or no data is being used in individual decision-making. And without data, decisions are based only on opinions. So essentially, the same people who often use data to make decisions in their professional life, are making decisions in their personal lives based on opinions.

I do wonder why. My view is that it is relatively easy to access data when it comes to organizations, because there are systems, processes and structures in place to access it. But when it comes to individual decision-making, accessing data has been a challenge. Making informed decisions often requires gathering and analysing data, but not everyone has easy access to the necessary information. There are multiple reasons for it such as limited access to resources like data subscriptions and the complexity of consolidating data from various dispersed sources.

More importantly, similar to how the lack of effective data culture is a challenge for organizations to make data-driven decisions, individuals are also not data-driven owing to the age-old way of

making decisions relying on gut and experience—our personal ‘HiPPO’. There are multiple reasons why despite having access to a wealth of information people often rely on opinions rather than data in their personal decision-making.

For example, emotional biases can exert a significant influence, particularly in emotionally charged situations, as people may rely on intuition or gut feelings over data-driven analysis. Then there are cognitive biases like confirmation bias or availability bias that can lead individuals to ignore or dismiss conflicting data. Additionally, the time and effort required to analyse data may discourage its use, favouring quick opinions. Lack of confidence in one’s ability to work with data and the bias of personal preference further support the inclination towards opinions. Moreover, the social aspect of seeking advice from others contributes to incorporating opinions into decision-making.

However, with the advancements in digital technologies like mobile devices, wearables and the internet, etc., the Big Data world has brought about such an abundance of data for individuals too, which provides them an opportunity like never before to incorporate data in their personal decision-making process as well.

But, before I start talking about benefits of data-driven decision-making, let me talk about why relying solely on gut for decision-making, although not a wrong approach, doesn’t always work.

Gut feeling and intuition

David Lewis, a renowned American philosopher said, ‘Our “intuitions” are simply opinions.² Often called gut feeling, it is also defined as a feeling that appears quickly in consciousness, with us unaware of the underlying reasons, but strong enough for us to act on as defined by Gerd Gigerenzer. Albert Einstein described intuition as ‘the highest form of intelligence’. But to me it’s a bit more than that. I define intuition as ‘a higher level of consciousness’, a capacity to sense energies and to create a ripple effect of wise choices.

In 2016, a series of experiments were conducted by the University of La Verne and Harvard University on employing one's gut feeling to accurately interpret other people's emotions. Results revealed that people often go wrong in understanding other's emotional state when relying on gut instinct alone. Participants across gender, age and education levels were found to be statistically inaccurate when they interpreted emotions of others solely based on limited facial cues. In contrast, those who employed systematic analysis of contextual cues, physical responses in addition to facial cues, demonstrated significantly better accuracy in understanding emotional states. This proves that while intuition is a response to unconscious processing of data, at times it results in sub-optimal decisions. This is because intuition is filled with numerous biases which makes it hard to differentiate an opinion from a real intuition.³

Therefore, it is advisable not to solely rely on gut instinct, rather look at data to make more informed decisions.

Consider replacing 'my gut feeling says to ...' with 'the data says ...' as much as possible

While I speak about intuition above, I feel that there is no clear path of developing the same. One possible way is to use data to back your decisions. It is increasingly happening in the corporate world, where data is used to make important decisions, but can it also be used to make individual decisions. A paradigm, when we can shift from saying—'I feel it in my gut' to 'the data says so'.

I have been impressed with the 'Thinking, Fast and Slow' concept of System 1 and System 2 thinking developed by Daniel Kahneman. While System 1 deals with instinct, System 2 deals with logic and information.

In life, we need both—sometimes System 1 and sometimes System 2. In his book, Daniel Kahneman, argues that System 1 is sometimes referred to as cognitive laziness or cognitive ease and making a decision only based on System 1 is prone to doubt, cognitive biases and errors, as it is used for instinctive decisions and

often relies on emotions or intuitions. Similarly, System 2 thinking is slow, deliberate and analytical. It involves conscious effort, attention and logical reasoning. And since System 2 thinking requires more effort and energy, people may tend to default to System 1. Learning to find the right balance between System 1 and System 2 is what we have to practice.

I would like to go deeper into this aspect of data replacing gut feeling at individual level by exploring the example of wearable devices for managing a lifestyle disease like diabetes.

For a long time, I was tracking my glucose levels with a prick in the morning every couple of days to track my 'fasting sugar'. Then three years back, I started tracking my glucose levels using Libre, a wearable sensor that continuously monitors body glucose levels. The impact has been nothing short of revolutionary. Despite being a data geek, I had seen blood sugar as a somewhat static number. Once I started using Libre and I could see the real-time data on a continuous basis, it was a shocker. Glucose levels are highly dynamic and sensitive. This understanding of trends and patterns helped my endocrinologist and me in making some important changes. The aha moment was when after analysing the pattern of my glucose levels, my endocrinologist decided to shift my medication timing from night to daytime. Why? Because the analysis revealed that my glucose levels were dropping during sleep and peaking late morning. Interestingly, 'stress' seemed to have a lot more impact on my sugar spikes than food—for example, during my son's squash matches! Moreover, I could track additional KPIs, like 'time in target' that helped me become even more cautious. Libre also gave alarms for both high and low glucose levels, allowing me to take immediate action.

Now, for years, my medication pattern that was similar to anyone else's was finally personalized only because I decided to go beyond just relying on experience or gut to relying on data, which my doctor was also confident to use to make more tailored decisions for me. It was no longer about how I felt, but about what the data said.

You don't need to be young (relatively!) and tech savvy to benefit from a wearable like Libre. It took months to convince my eighty-

year-old father-in-law to start using it. Since he has started using it, it has made a world of difference. With him, it started with denial. According to him, he hardly ate any sweets (many of us claim that!) and his blood glucose levels should therefore be well in check. So, we made him monitor his sugar levels on his Libre after every item that passed his lips. In little over two weeks, he was able to see the patterns around his blood sugar levels and his food intake, his activity levels and especially the havoc that sweets intake caused for him. He now wears it religiously and it monitors the peaks and drops of his blood sugar every hour, makes them into trends and patterns and matches that to his body pattern.

Instead of relying on his 'gut' (which he realized was often wrong), he now has solid data to look at and make informed choices, replacing the food he is eating or changing the time of a meal. As a result, he has cut out sweets completely and has begun to walk regularly. In little over two weeks, he developed a data-driven approach to zero in on the best course of action for his body. A behaviour that we as a family failed to exert any influence on for over a decade, changed in just two weeks thanks to a wearable like Libre which is easily available in the market today. My father-in-law has become more mindful of data and now utilizes it to validate and, more frequently, correct his instincts or 'gut feelings'.

In fact, an overwhelming 86 per cent of surveyed patients reported that their devices had a positive impact on their health and overall quality of life, while also enabling their doctors to deliver a higher standard of care.⁴ If you are looking to build new habits or manage your energy, the outcome can be easily achieved with the wearable technology available. Today, data removes ambiguity from decision-making, even at an individual level, and helps us make decisions for living a healthier life on a day-to-day basis. No wonder, the wearables market is growing at 18 per cent CAGR, suggesting that the desire of individuals to monitor themselves is high.⁵

Importance of data-driven decision-making

Why is this important? After all, we have lived a number of years without data-driven decision-making, then why do I advocate this so strongly?

Data-driven decision-making improves the quality of decisions and reduces the chance of errors. It ensures that your biases are made visible, and even after that, if you choose to make a decision around 'what feels right' vs what the 'data says', it is a conscious choice you make.

I believe that adding data into individual decision-making helps will help each person to live healthier, like I illustrated with the wearables example; plan better, like planning a trip; save better, by securing one's future through sound financial planning; and work better, by setting clear goals and using data driven tools for better decision-making and greater impact at work.

Integrating data into your personal life can be as simple as the data collected by your smartphone about your activities or leveraging the plethora of sources available to you. Here is an example of how your ability to understand and apply the valuable information derived from this data can help you plan better.

Let me ask you this, when was the last time you dined out, stepping out of your house without scanning for options on your phone? I bet you didn't, because we are living in a world of online searches, booking and reviews. Crowd-sourced reviews and social networking apps like Yelp, have fundamentally changed the dining experience for every person around the globe. So before grabbing the car keys, people categorize restaurants first by the cuisine they are in the mood for. Then by ratings, time taken to reach and so on. They then look at the reviews and the recommended cuisines. Apart from that there are multiple other factors like kid-friendliness, band playing, open area that a person can filter for. And after such extensive evaluation, a restaurant is selected, online booking is made and that's when you grab your coat and your near and dear ones to enjoy a pleasant meal, where you have carefully chosen almost each and every aspect of your dining experience. And yes, it

makes a significant business impact too—a HBS study declares that just a one-star increase in Yelp rating by customers leads to a 5–9 per cent increase in restaurant's revenues.⁶

DIAI framework to help you in data-driven decision-making

Now that we know what is possible, let's look at how the use of data can enable individuals to derive greater value at every stage of the DIAI framework that I talked about in Chapter 3, Value Reimagined.

Let me illustrate with an example here to make it relatable and easier to understand. Let's say you want to use data to make better financial decisions. So, let's break down how the DIAI framework can make a real difference when it comes to making smart financial investments:

- **Data:** Multiple data points related to your own financial situation, such as your income, expenses and any investments you already have will help you understand your inflow and outflow and understand your investment potential and risk appetite. Combining your data on your personal situation with data on varied investment options that are available in the market, data on market trends and economic news that might affect your investments helps you make informed choices on the type of investments that best suit your needs.
- **Insight:** Analysing various investment options to uncover trends, either through your own analysis or the analytical tools available in the market today, like portfolio management tools, or share market analysis tools, etc., can help you generate insights like spotting patterns, such as investments that are doing well lately, or those that are expected to perform well in future and what kind of returns can you realistically expect from

different options? You also evaluate the level of risk involved with each investment option.

- **Action:** Based on the insights generated, you are able to create a plan considering your financial goals, how much risk you're willing to take and how long you plan to invest. You also choose your asset allocation, where to put your money. Stocks, bonds or maybe a mix of both, that best reduces your risk exposure.
- **Impact:** Once investments are made based on data-driven decisions, the impact stage represents the period during which the anticipated target returns are realized. It's the phase where the effectiveness of the investment strategy is assessed, and adjustments may be made to optimize performance and ensure the desired outcomes are achieved.

By following this framework, you can make investment decisions that are backed by data, to help you stay focused on your goals and minimize the impact of emotional reactions. Over time, it can lead to more successful and informed financial choices tailored to your unique situation.

The DIAI framework works for everyone, irrespective of who they are and what they do. For example, it can significantly enhance the value from data even for a farmer! See how the various stages of the DIAI can enhance value for one:

The world of Big Data provides farmers granular data on rainfall patterns, water cycles, fertilizer requirements and more. This enables them to make smart decisions, such as what crops to plant for better profitability and when to harvest. The right decisions ultimately improve farm yields.

As the world population is on a rise, there is an urgent need to increase food production. Global population will reach 9.7 billion by 2050,⁷ a 1.7 billion increase from now,⁸ and with climate change, rapid urbanization, there is a dip in the total land under farmlands, and hence the food production. On the other hand, a third of the

food produced is wasted. It is therefore important that we get better at crop production and crop management.

DIAI framework for better farming



This is exactly what IBM is helping farmers achieve. Its Watson Decision Platform for Agriculture combines data points from satellites, drones, weather models and IoT sensors using IBM PAIRS Geoscope, that provides easy access to useful insights and farming tips for better crop management. With the help of this AI-enabled tool that analyses the vast amount of data generated (set to reach 4 million data points per day by 2036), a farmer can significantly enhance the crop yield and manage it better by identifying best practices, improving crop yield, managing irrigation, assessing pest and disease risks, and comparing field conditions. Additionally, the IBM Weather Signals also helps farmers predict consumer reactions to weather in real-time which helps them plan production, logistics and supply chain accordingly. Availability of tailored models for various crops and geographies benefits both growers and the broader agricultural ecosystem.⁹

Similarly, InVivo, France's leading agricultural cooperative group with 220 members and €6.4 billion in sales, has developed an app, SMAG, which has thirty years' worth of weather data, satellite and drone images and soil types—to make informed decisions faster. Again, making the data accessible to farmers in the form of an app, helps deliver impact and make decisions on the go. It is being used

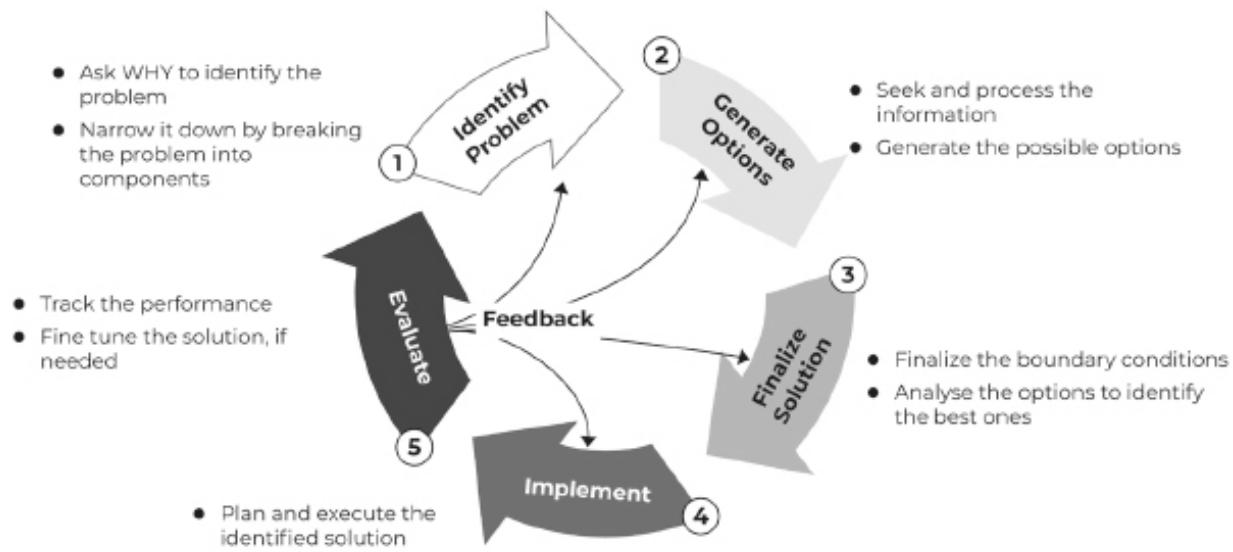
by 80 per cent of cooperatives and 50 per cent of merchants in France.¹⁰

Getting better at applying the DIAI framework for individual decision-making

While the framework is rather intuitive, one does wonder, how does it actually apply to self. How do we implement this framework in our individual life, when we are stuck in any situation. What do you think? How to break down a problem? What questions to ask?

To begin with, let us acknowledge that it is not a linear process. It is a continuous and iterative process, which when done repeatedly enables us to get better at it each time. But the process predominantly can be broken down into five simple steps that every individual can take, when they are trying to make a decision. It is noteworthy that every stage of the five-step problem-solving is anchored in data.

Five-step data-driven problem solving for individuals



As individuals assess the impact of their decisions, their proficiency in the preceding five steps grows, leading to the **generation of greater value**

Identifying the problem: In today's fast-paced and constantly connected world, it's easy to become overwhelmed and lose sight of what truly needs our focus. Thus, it is essential to take a deliberate step back from the chaos and **engage in a reflective process**. To introspect and think deeply about the problem can help get rid of the complexity and break the problem down, narrowing it to the core issue. Data plays a key role in the reflective process ensuring that introspection and problem analysis are based on actual facts rather than assumptions or perceptions. As a result, we can gain a clearer perspective on the issue. For example, let's say you're facing persistent fatigue and low-energy levels in your personal life. You step back, introspect and reflect on your lifestyle and identify that the root cause is your sedentary lifestyle and lack of physical activity. To gain insights into the issue, you may collect data related to your lifestyle, such as your daily routines, sleep patterns, diet and exercise habits.

Generate options: The democratized access to information can help identify multiple options for problem-solving. So, at this stage, one has to broaden their horizon, to make sure one looks at varying perspectives, to gain a multidimensional view of the possible options available to solve it. By actively **seeking diverse perspectives**—online, offline, friends, relatives, experts, etc.—one is able to accumulate a multitude of ideas and approaches that they might not have considered otherwise. The wider one goes, the pool of potential options enriches, offering a comprehensive spectrum of choices to explore. Here, data serves as a critical tool in the process of broadening one's horizon and generating a multitude of options for problem-solving. It ensures that options are well-informed, evidence-based and inclusive of diverse perspectives.

For instance, to restore physical health, you recognize the need to make changes. You access online content, consult friends, health professionals and self-help resources from various backgrounds. Suggestions range from time-management techniques to health supplements to incorporating exercise and meditation into your routine. You collect varying data points like evidence-based

information from health websites and platforms, research findings, anecdotal data from personal experiences, clinical data, and test results to generate feasible options.

Identify the best option: At this stage it's crucial to acknowledge that every decision comes with its own cons. Therefore, it is important to meticulously map out the **pros and cons** of each option on the table. This step allows you to weigh the benefits against the drawbacks and gain a clearer understanding of the trade-offs involved in each decision. Once the pros and cons have been thoroughly examined, the focus shifts to selecting the **best option** that offers the most optimal solution to the problem at hand. Data allows for the quantification of the potential benefits and drawbacks, analyse the risks associated with each option, compare with historical performance, and explore 'what if' scenarios. You do the same with the list of options you produced to gain better health. You do a thorough assessment and evaluate the pros and cons of each option. For example, exercise and meditation may require time and effort for enrolling and going to classes, but they offer long-term health benefits. You compare the financial investments required, the time commitment and measure against the long-term health benefits (stamina, strength, stress level, etc.) to narrow down the most suitable approach for you.

Implementing: While implementing the best possible solution, it is important to adopt a broader approach which entails considering **all aspects of the solution** that includes the necessary resources required and the process to effective implementation. When dealing with significant changes, it is advisable to progress towards the ultimate goal through **small, incremental changes**. Data can inform resource allocation decisions by providing insights into the costs and resource requirements associated with each step of the implementation process. However, patience is key during this phase as it is important to stay the course and maintain a consistent effort. Say you decided that the best option is to meditate and exercise. You then decide how to implement it. Which classes to join, whom

to follow, what books to read, etc.? Given this has to become a way of life, you start small, maybe enrolling in a yoga class and then gradually add variations to it like meditation or aerobics.

Evaluation: The evaluation phase in decision-making is a pivotal step that involves reflecting on the consequences of your choices, both positive and negative. It's an opportunity to **assess the impact** of your decisions on your goals and you, acknowledge your successes or identify areas for improvement and foster a **mindset of continuous learning and adaptation**. Data can help track progress against your goal, monitor milestones and create a feedback loop to capture the effectiveness of each incremental change. As you continue working on your physical health goal, you regularly evaluate its impact on your energy levels, well-being and overall happiness. You recognize that some aspects are working exceptionally well, such as increased energy and reduced stress. However, you also identify areas for improvement like needing to change your diet or mix-matching it with kickboxing or completely replacing it with something else that you enjoy more like dancing. One basic way is to measure your health metrics like weight, body composition, blood pressure and fitness levels, to objectively assess the impact of your lifestyle changes. Based on the evaluation you can decide what more or less you can do to stay on track.

As you have probably already sensed, this is an iterative process, and that's the essence of getting better at it. As you assess the impact of your choices, you gain proficiency in identifying the right problem the next time, generating more effective options, selecting the best-suited solution and executing it with improved planning. Over time, you'll notice that each cycle of data-driven decision-making becomes smoother and more efficient, easily integrating into your decision-making routine. Ultimately, it evolves into second nature, enhancing the quality and precision of your decision-making process.

Over the years, data has become abundant and democratized. It is now possible to improve our lives and make better decisions using data. We can be more informed about the choices we are making.

However, we should not ignore the value of intuition. An individual, after going through the five-step framework, may still choose to go with a decision that is not anchored in data—but that awareness and the option to take a decision that is anchored in data is an important aspect which helps each one of us know ourselves better. If we keep track of this, maybe, over time, this will start informing our intuition and make us better intuitive decision-makers as well.

Looking into the future, I will hazard a guess that people who follow the five-step method of decision-making, people who start tracking their decisions anchored in data, may land up developing a stronger sense of intuitive decision-making, as data-processing will become second nature to them. Like anything else, this is also about discipline and practice, in a world that has abundant data to pore over.

Key takeaways

- Despite having access to a wealth of information in the digital age, people continue to favour opinions over data in personal decisions due to emotional and cognitive biases, quick convenience, limited data analysis skills, personal preferences and social influence.
- Intuition is a rapid, subconscious response, often without awareness of the underlying reasons, and hence can be filled with several biases, sometimes blurring the line between genuine intuition and biased opinions, making it a valuable yet imperfect tool for decision-making.
- Data-driven decision-making reduces ambiguity, helps avoid major errors, exposes biases and empowers individuals to make informed choices based on factual information, enabling individuals to bring positive impact in every aspect of their lives.
- The DIAI framework is an effective way for individuals to derive value from data. For effective implementation of the DIAI framework, individuals must follow an iterative,

five-step process, beginning with identifying the core problem, gathering information, exploring solution options, executing the chosen solution and evaluating performance.

23

Information and Wisdom

Reflect and Recognize Patterns

'Where is the Life we have lost in living? Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?'

—T.S. Elliot,
Poet, essayist and literary critic

Wisdom as a concept has captivated the imagination of thinkers and philosophers over the years across civilizations. It has an ancient history, tracing back to Sumerians around 2500 BCE, where 'Nanna' embodied the god of moon and wisdom.¹ In Hindu mythology, Lord Ganesh, widely regarded as the god of wisdom, is said to have written the Mahabharata, an epic known for its treasure trove of wisdom.

You might wonder why I am talking about wisdom in a book on data. In the last chapter, I talked in detail about the importance of data-driven decision-making for individuals. But data-driven decision-making isn't sufficient on its own, and I can say this from my own experience of working with some of the leading organizations in the world. These organizations have access to vast amounts of data but still make mistakes when they don't think deeply. Relying solely on data and information can give a false sense of confidence and a superficial understanding that some people confuse with being wise.

This is where the paradox of information versus wisdom comes in. With the advent of Big Data, we have access to a lot of information but does having access to a lot of information make us wiser, or does it just make us feel wiser? Does data abundance enhance or

hinder wisdom? One might think that the importance of wisdom has gone down in the Big Data world but, to my mind, it has become even more crucial. Wisdom is like proprietary knowledge that makes people stand out, helps them make smart choices and be more effective with the data-driven decisions that they make. The interplay between information and wisdom echoes the enterprise dynamics of multi-source data and proprietary knowledge.

Gaining wisdom has never been easy. It involves years of learning, experiencing and reflecting. In today's Big Data world, data, in abundance, has the potential to help build wisdom in numerous ways. It helps us make informed decisions, uncover patterns and learn continuously. It also allows us to reflect on past decisions and adapt for the future. However, the Big Data world comes with its own unique challenges, complexities and paradoxes, such as information overload, digital echo-chambers, short attention span and so on.

But the good news is, if leveraged well, data can surely enhance the process of building wisdom. At the same time, I would like to highlight that the process of building wisdom requires time and experience, which people often ignore in the Big Data, high-speed world we operate in today. In this chapter, I will dive into the idea of wisdom and talk about practical ways to nurture it in our modern, data-driven world.

Before we get to the process of building wisdom. Natural question that comes to mind is 'what is wisdom?' This is a difficult question to answer. Let me explain why.

Why is defining wisdom difficult?

Over the years, numerous attempts have been made to define wisdom, highlighting a different perspective on wisdom each time. There is the cognition-focused definition anchored in critical thinking, sound judgement, knowledge, experience and insight, emphasizing the pursuit of a deeper life meaning and optimal living. There is another one, a personality-focused definition, that focuses on traits like compassion, reflection, open-mindedness and humility as key

markers of wisdom. Yet another one is development-focused and defines wisdom as a product of life experiences and personal growth, emphasizing that true wisdom emerges not solely from experiences but from active reflection and the ability to derive meaning from life's lessons.

Despite the multiple definitions available out there, there isn't a single definition that prominent thinkers and philosophers agree on. Why is it so hard to define wisdom? Firstly, it is a multifaceted concept, with different disciplines offering different definitions based on specific context. But I believe the core reason behind it is that wisdom cannot be measured. There are no clear metrics (KPIs) that can be used to measure how wise a person is. It cannot be measured as physical attributes like height or weight. Rather, it is a bit like attempting to quantify intangible qualities such as happiness or love.

But despite these challenges, for me, there are three underlying aspects that are core to wisdom. First is **thinking deeply about things**. It refers to engaging in thorough contemplation or reflection on a particular subject or topic. It requires going beyond the surface, analysing the complexities and considering various aspects or perspectives of the matter at hand, both inputs and possible outcomes. It also implies going beyond the short-term and thinking through the long-term implications. The second one is **upholding strong values**, which means anchoring on fairness and sustainable outcomes in various situations and decisions. These two put together result in the ability to make **sound judgement**. Sound judgement enables individuals to make reasonable and well-informed decisions based on careful consideration of multiple facts, evidence and relevant factors, that are not clouded by ego and short-term self-interest. In summary, wisdom is the ability to assess situations deeply, think about the pros and cons of different options objectively, think about the long-term impact of decisions and make choices that are sustainable and in the best interest of the people and situations involved.

I have met many such wise souls over the years, and they have shown me that wisdom can take any form. A personal experience

from my Fidelity days is a testimony to that.

During those days, I recall partnering with a senior lady, who was twenty-five years my senior, a counsellor to the chairman at Fidelity Investments. I was leading an initiative of global significance and was using my own unique approach (which was not in line with the traditional Fidelity approach). I faced a lot of opposition from the senior executives on my recommendations and was told that I would lose my job if I went ahead. However, very surprisingly my presentation to the chairman went very well and he agreed with my aggressive recommendations. I later realized that it was the 'wise counsellor' who had been closely observing me at work who had briefed the chairman beforehand! I then saw the value of this group of trusted counsellors of the chairman, who had probably led, managed and seen many such strategic initiatives. This lady could understand the worth of my ideas, even in the brevity of a few meetings, and to give me a go-ahead, even if it was a maverick idea. The ability to make a decision, which possibly has long-term implications, takes a lot of clarity, belief and courage. The decision was coming from a place of wisdom. It was evident.

When we think of wise people, we think of people we want to reach out to for guidance. When we think of wise decisions, we think of situations that will have a long-term impact that is often hard to understand in the current context. Operating from wisdom often feels like operating from a crystal ball; when you are able to see the unseen!

In fact, wise people are also more effective in leveraging data for their decision-making. They recognize that wisdom isn't about disregarding data but rather complementing it with their wealth of experience and insight. Let me explain the critical role wisdom plays in the data-first world.

Importance of wisdom in data-first world

The abundance of data does not necessarily mean better decision-making for individuals. As we talked about in Chapter 22, Data for Better Decision-Making, making data-driven decisions requires us to

go through a five-step process: identifying the problem, generating options, finalizing a solution, implementing it and evaluating the impact. Identifying problems and generating options are often not fully comprehensive, requiring using our wisdom and reflecting. Similarly, evaluating the impact also involves judgement and reflection. As you can see, our wisdom has a significant impact on our decision-making process. It assists in evaluating the long-term outcomes of our actions concerning the problems we choose to address and the solutions we decide to implement.

Typically, as we have data, we often have a false sense of belief that we understand the problem, when in fact, we have often not even scratched the surface and we do not fully consider the consequences and see beyond what meets the eye. Wisdom plays a critical role to be fully aware of these consequences. Many people, when working with one set of data will come up with some insights and with another set of data, will come with different insights. Their ability of data-interpretation is high, but pattern recognition to map both the data sets may be low. Wisdom has to do with pattern-recognition. In recent times, we have seen the fall of some of the biggest banks in the United States. In March 2023, the \$200 billion Silicon Valley Bank declared bankruptcy—as they had racked up losses due to poor investment decisions. In a transaction, they had raised \$1.8 billion to cover losses.² As the news went public, people started withdrawing and it all went downhill from there. The situation was similar to what happened in the Great Depression of 1929 and the Global Financial Crisis of 2008. Clearly, there were poor decisions at multiple levels in large financial institutions that led to such catastrophic failures. These financial institutions possess extensive data resources and a talented workforce. I often wonder how highly paid, experienced executives who have abundant access to data and analytics end up making decisions which had such severe repercussion? Is that a lack of wisdom? A lack of long-term thinking and an inability to think through the consequences of decisions?

We really do not know and we have no way to test it. But I do believe, just data analyses cannot give you all the right answers. In a data driven world, we get a false sense of security by just holding on to information, we believe we know, and we have assimilated, when in truth, we have just stored the information. We are just consumers of information. It is wisdom that helps us make the shift from consumers of information to users of information.

I am not saying that wisdom and information are disconnected. Rather, wisdom plays a crucial role in driving value from information. Let's look at the connection between data, information and wisdom to understand this better.

A framework for wisdom

The most widely accepted model, of the modern times, to understand wisdom is the DIKW hierarchy also known as the wisdom hierarchy, knowledge hierarchy, information hierarchy, information pyramid or the data pyramid. This model representing data, information, knowledge and wisdom, gained attention when Russell Ackoff discussed it in 1989,³ but its roots trace back to T.S. Eliot's 1934 poem 'The Rock',⁴ where he pondered the loss of life in living, wisdom in knowledge and knowledge in information. The idea resurfaced independently in the late 1980s as concerns about information-overload grew.

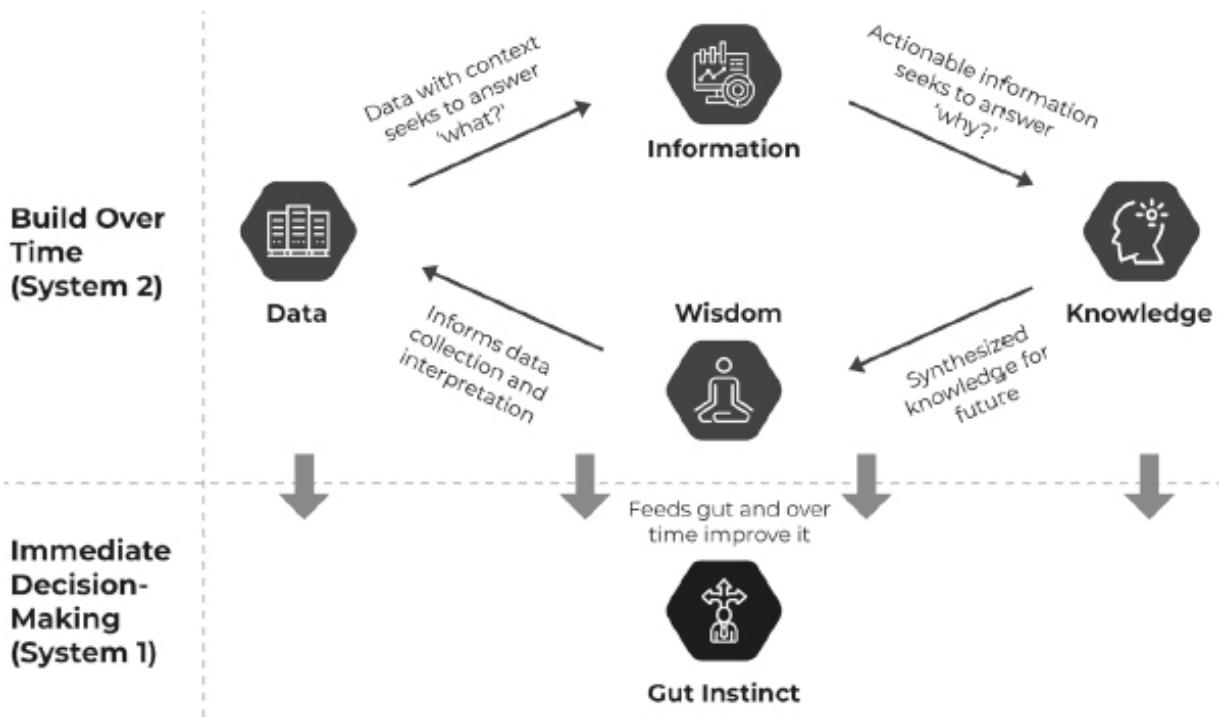
The DIKW model proposes a step-by-step progression: data generates information with context and structure, information leads to knowledge and knowledge eventually results in wisdom. Data, in its raw form, is the unprocessed collection of facts, figures and observations. It is like having a list of numbers without any meaning. Information transforms this data by organizing, structuring and adding context, answering questions like 'what?', 'who?', 'where?' and 'when?'. Moving up the ladder, knowledge takes the understanding a step further by delving into the 'why' and 'how' aspects, analysing patterns and gaining insights from the information. Wisdom, the pinnacle of the DIKW hierarchy, involves

not only understanding but also the ability to make sound judgements and ethical decisions based on knowledge and experience.

While I appreciate the elements and flow of the DIKW hierarchy, I believe there are some inherent issues with this model. The concept of a pyramid with a linear, one-way progression and its static nature does not look like a fair representation of the path to wisdom. In the real-world scenario, the path to wisdom is not that straightforward. It is much more complex. In my view, it often involves iterative and cyclical processes of refinement and continuous learning.

Envisioning this concept as more of a continuum rather than a rigid pyramid makes more sense to me. In this fluid continuum, raw data evolves into information when paired with 'what' and structural context, providing a foundation of understanding. As we delve deeper, this information transforms into knowledge, where we not only possess answers to the 'why' but also actionable insights, signifying a deeper level of comprehension. Further, along the continuum, as we synthesize this knowledge, we start to identify patterns, based on which we make decisions that shape our future—this is where wisdom resides.

Wisdom and gut in a data-first world



Wisdom is not built in a day. It is built over a period of time. So, as we further continue in this loop, wisdom acts as the guiding light, making us better at navigating the overwhelming deluge of data that surrounds us in the digital age. It assists us to filter out the irrelevant data that can overwhelm us and instead focus on what truly matters. With this refined ability to distinguish, we rise above the noise and make choices that align with our long-term goals, values and the betterment of the world and ourselves.

As you can see in the framework above, wisdom and gut are also interlinked. Over time, as we built wisdom, this wisdom feeds into our gut instinct, making it sharper and more accurate. Let me explain how.

Cultivating a 'wise' intuition

Although the basic ingredients of wisdom and gut are data, information and knowledge (DIK), wisdom is built through deep synthesis of these ingredients. It involves thoughtful processing and

drawing upon one's experience and understanding. It encompasses a holistic perspective, making it comprehensive and well-informed.

Whereas gut instinct represents an immediate and instantaneous response. It is an intuitive and quick reaction to a situation or problem. The basic ingredients of gut instinct and wisdom are the same but gut instinct does not involve deep processing of DIK, often relying on first impressions and immediate reactions.

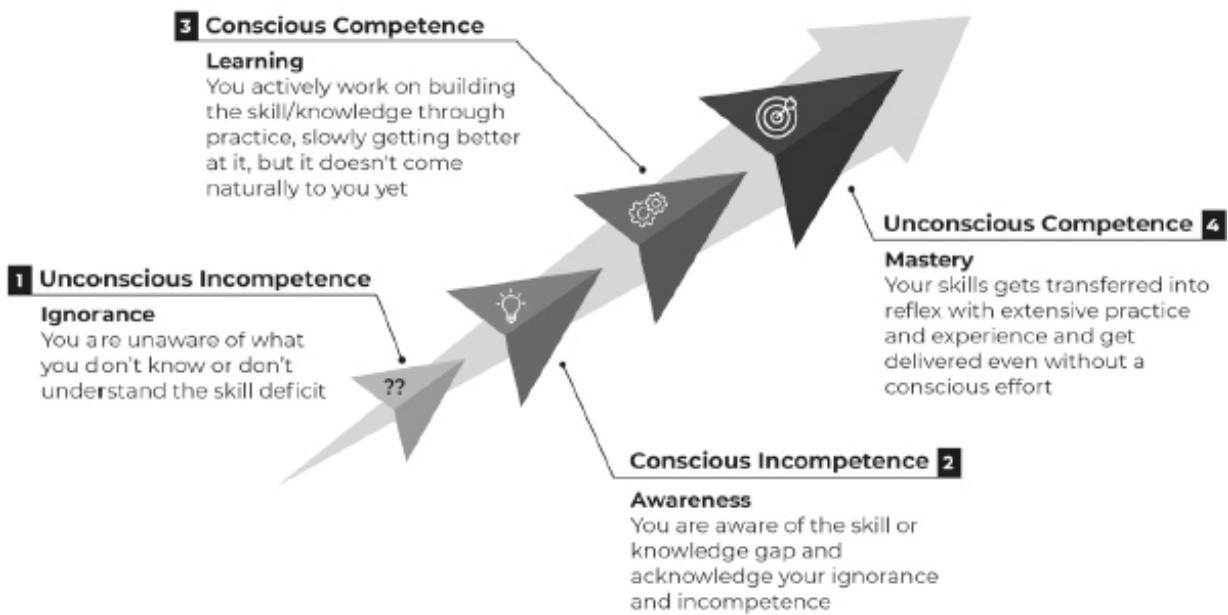
In the previous chapter, I have already talked about System 1 and System 2 thinking for decision-making. Now, let's explore how these systems interact to influence the relationship between gut instincts and wisdom.

Initially, any complex decisions require a deliberate and thoughtful process, referred to as System 2 thinking, which involves going through the entire DIK cycle, starting with data, then analysing data, considering various factors and making choices based on logic and reasoning. It is a conscious and systematic approach to decision-making which in the long term helps build wisdom.⁵

However, with time and experience, our decision-making process evolves. Repetitive exposure to similar scenarios, transforms this once conscious process into something more automatic, subconscious. It becomes an intuitive part of our decision-making, referred to as System 1 or gut instinct. Our gut instincts, informed by our learnings, experiences and wisdom, prove invaluable. It is as if our intuition has become finely tuned, allowing us to make swift, yet well-informed decisions.

It's crucial to note that gut instincts are not always right. They vary from person to person and across different domains of knowledge and situations. For instance, I can confidently say that I have built a highly effective gut instinct in matters related to people, problem-solving and getting to the core of issues. Because these are the areas where I have dedicated years of practice and learning. On the other hand, when it comes to stock-market investing, I would prefer not to rely on my gut instinct, as I do not have deep experience in this space.

Four stages of competence



Why are some gut instincts very accurate while some are highly prone to errors and biases? Let me talk about the competence framework, where the four stages of this framework work well in understanding how gut instinct is built through experience, learnings and wisdom. ⁶

First, there is the **unconscious incompetence** stage, where we do not even realize what we don't know. At this point, we might believe that our gut instinct is right, but our gut instincts are not reliable, and decisions based on them might be wrong. Then comes **conscious incompetence**, where we recognize our lack of expertise in a certain area, and we learn not to trust your gut instincts there. Instead, we rely on data and knowledge. The third stage is **conscious competence**, where we have developed skills and understanding. We can make informed decisions, but it takes effort and conscious thought. As we repeat this process, it helps us identify the patterns and reflect on our learnings. Finally, we reach **unconscious competence**, where our capabilities become second nature. In this stage, our refined gut instincts, shaped by learning, experience and wisdom, become highly effective and trustworthy for decision-making. It might look like the person has a strong gut

instinct and it's right most of the time, but it is actually the outcome of years of practice and accumulated wisdom.

A classic example of unconscious competence can be seen in the world of sports, such as with tennis legends Roger Federer and Novak Djokovic. Aren't we all amazed how they are able to anticipate the movement of the ball and react within split seconds. It's nothing but their gut instincts operating at the level of unconscious competence. Their skills have become so ingrained through years of practice that they can deliver remarkable performances without conscious effort.

There are two significant interplays here: one between information and wisdom, and the other between gut instinct and wisdom. Information is a key input for building wisdom. Deep reflection on information and experiences helps us build wisdom. At the same time, wisdom helps us filter out the right information from the sea of information we have access to in the Big Data world, which, in turn, aids in building wisdom. The interplay between gut and wisdom is intricate; both draw upon data, information and knowledge as input, but the output differs significantly. Wisdom is more reflective and long-term, whereas gut is more instantaneous and immediate. The gut also goes through stages of transformation, from highly ineffective unconscious incompetence to highly effective conscious competence. These interplays remind us that wisdom is not static; it's always evolving. It grows stronger as we learn, gain experience and build reliable instincts, helping us make smarter decisions in the ever-changing data-first world.

Is wisdom dying in the data-first world?

In today's world of instant gratification and data abundance, where we are all bombarded with information, and consuming data like never before, finding time for reflection is tough. But wisdom relies on reflection. I often wonder: is wisdom dying?

As much as I would like to say no, observing the changing world around me, it seems like wisdom is indeed fading in our data-rich era because we have forgotten to step back and reflect. In the age

of information overload and countless digital distractions, it is very easy to lose sight of what truly matters. We often prioritize quantity over quality, speed over depth and noise over contemplation. The constant notifications, social media updates and the glamour of viral content draw our attention away from the meaningful and towards the insignificant. As we engage with these digital distractions, we inadvertently distance ourselves from the deeper insights and understanding that wisdom offers.

As data explodes, there is rising FOMO (fear of missing out) and more and more information is consumed. Every individual stores the information they find relevant, without really reading it—it is indexed, but not converted to knowledge or wisdom. Depth of knowledge on subjects is on a decline, quality of conversations is on a decline, but volume is on a rise. People are confusing wisdom with information or gut feeling.

Wisdom is not just information; it is informed by data and information. Wisdom is not gut—it is more than an instinct; it is not just instantaneous but reflective. It is insights meets knowledge meets experience. Only those who take the time to pause, reflect and truly absorb knowledge will become ‘wise men and women’.

Wisdom, as we are today, is dying. So, is there a way to save this endangered but critical facet of human understanding?

Five principles for building wisdom

Can wisdom be cultivated? I believe that the answer to this question is a resounding yes. Absolutely! Wisdom can be nurtured and cultivated. However, it’s important to understand that it’s not a one-time effort. Instead, it’s an ongoing journey of learning and reflecting on experiences to improve decision-making in all aspects of life.

When I think about wisdom, for me, it goes beyond simple insights, requiring a profound and well-rounded comprehension that can hold up to scrutiny from different angles. This clarity is achieved through thorough study and understanding. Take Warren Buffett, for example. He is a wise investor because he relies on his principles,

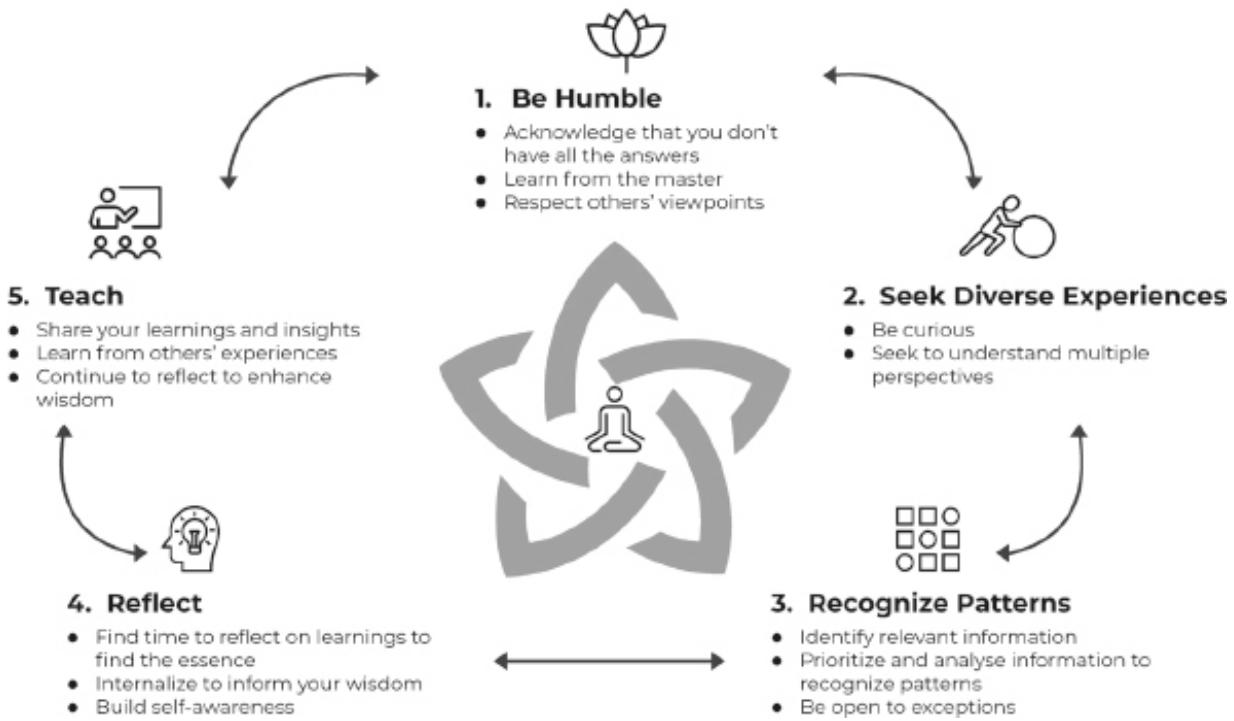
built through deep learning and experience. He invests for the long term, often visiting companies he invests in to solidify his understanding. Buffett's beliefs may differ from the norm, but they are shaped by his unique interpretation of data, experience and past bets. His ability to connect dots, identify patterns and refine his approach over time exemplifies wisdom.

I consider myself a student on the path to wisdom, a journey that demands reflection and continuous learning. Reflecting on the insights of experts, I have come to realize that wisdom often begins with quiet moments of introspection, where the five principles I will discuss below gradually become habits.

Each of the five interconnected principles contributes to developing wisdom. These components of wisdom all connect and work together, like a beautiful puzzle coming together. Let me explain each of these principles in brief and how I have applied them in my professional and personal journey.

Let me start with **humility or being humble**. It is about self-awareness, acknowledging what you do not know, listening without judgement and being open to others' ideas. In essence, it is about being receptive to different perspectives and being willing to consider them with an open mind. As you grow and succeed, the same success can easily turn your head, making you self-absorbed and egotistical. And with that, one stops listening and being dismissive of others' viewpoint. So, I make conscious efforts to be more inclusive in my conversations and endeavours and seek diverse perspectives as much as possible, which keeps me grounded.

Five-part framework for building wisdom



Next comes **seeking diverse experiences**. Think of it as being curious like a child, exploring outside your comfort zone and learning from others. When you do this, you gain a broader understanding of the world and can make wiser decisions because you have a wider range of experiences to draw from. For me, I continually challenge myself with different pursuits like playing sports, meditation, travelling and meeting new people. Trying something new not only introduces me to fresh knowledge but also deepens my understanding on varied topics. I try to learn from the masters, as much as I can, which is both humbling and also offers me deeper and richer experiences.

Moving on, when we have gathered all this information and started organizing it effectively, we do start **recognizing patterns**. This means we begin to see connections and similarities between different pieces of information. It is like solving a puzzle where you start noticing how different parts fit together. Recognizing patterns is crucial because it helps us understand things at a deeper level and make more informed decisions. Merely accumulating diverse

experiences is not enough. They are only relevant when you derive the ‘so what’ from them. It helps in connecting the dots and recognizing the repeatable patterns that are crucial to develop frameworks or approaches to effectively deal with recurring situations.

But wisdom isn’t complete without **reflection**, another very crucial component. In our busy lives, we often forget to pause and introspect. Reflection allows us to make sense of our experiences and gain valuable insights, which are essential for recognizing patterns and developing wisdom. Taking time to reflect deeply is crucial. With the constant flurry of activities, there is a lot of clutter around me. So, I make sure I take time out for reflection, stepping away at regular intervals. Because with reflection comes clarity—clarity of thought and action. I have dedicated a separate section to it, ahead in the chapter, because its significance cannot be overstated.

Finally, wisdom is not about keeping what you know to yourself. It is about **teaching and sharing** what you have learned with others. Instead of hoarding knowledge, you spread it far and wide. I am passionate about teaching and sharing the knowledge I have accumulated over the years. It is one of the reasons I am writing this book. I believe that knowledge is that one thing that grows through sharing. And as I share, I get an opportunity to test my knowledge, creating a feedback loop, which is critical to keep learning and growing.

So, the five wisdom principles—humility, seeking diverse experiences, recognizing patterns, reflection and sharing—are like interconnected threads. They weave together, with each principle influencing and enriching the others. For example, when you approach life with humility, it opens you up to diverse experiences, which, in turn, can help you recognize patterns more effectively. These patterns, when reflected upon, deepen your insights and the knowledge gained can be shared to benefit others. Teaching others can be a humbling experience because it often reveals areas where you might lack knowledge or need further learning. When you share what you know, you might encounter questions or perspectives that

highlight your own gaps in understanding. This realization keeps you humble, reminding you that there is always more to learn and discover. Likewise, starting with any of these principles can lead to a continuous cycle of wisdom development.

Wisdom is not a destination; it is an ongoing journey of learning and growth. Much like the process outlined in my framework for building wisdom. It's similar to any skill; practice makes you better over time. Wisdom is not a one-time achievement; it is a continuous refinement of your ability to make sound judgments and navigate life's complexities.

Find time for reflection

Reflection is a key element for attaining wisdom. In today's fast-paced world, we often do not have time for reflection or self-reflection. But true wisdom, especially in leadership, comes when individuals look within themselves, take responsibility for what they do and embrace personal growth. Wise leaders not only analyse themselves but also are willing to adapt their beliefs and behaviours. If someone is in a senior role and there are problems in the organization, they likely played a part in causing these issues. Acknowledging this is the first step towards solving these problems for the betterment of all. This underscores the importance of self-awareness and acknowledging how they have contributed to the current situation and making deliberate changes for the greater benefit. Self-awareness, which involves understanding one's values, motivations and influence on others, lies at the heart of this process. Numerous academic definitions of wisdom are grounded in the principles of self-awareness and self-reflection.⁷

In a data-first world, the role of personal insight, intuition and judgement retain their significance. The path to nurturing these qualities begins with establishing a deeper connection with oneself. This helps sharpen your pattern recognition, which lies at the core of wisdom. In today's world, where data alone is not enough to stand out, your wisdom truly makes you unique.

Reflection, when rooted in self-awareness, fosters wisdom. For me, meditation is a potent tool for connecting with your inner self. This practice delves deep into self-exploration and consciousness, to connect with your inner thoughts and emotions. By quieting external distractions, it helps you gain insights into your values, motivations and inner processes, ultimately enhancing your self-awareness. This heightened self-awareness, combined with acknowledging your role in the current situation, is a crucial step towards wisdom. I will take this topic of meditation in detail in Chapter 25, Digital Engagement vs Mental Health.

Harvesting wisdom: From the individual to teams to organizations

Wisdom of teams is greater than the wisdom of individuals. The diversity of a team helps shape the team wisdom, each of the team members brings their 'what'—which collectively develops into a 'why' and 'how'—which helps in creating team wisdom. In Introduction, I spoke about the idea of 'wisdom of crowd', which has helped develop some of the best products and services in the world. For instance, Google's PageRank algorithm is a powerful solution created by leveraging 'wisdom of crowds'. The algorithm looks at how many links refer to a webpage and how good those links are. These links act like votes or collective 'wisdom of crowds', showing which webpages are more important. Pages with better and higher number of links, get ranked higher in search results. This enhances the quality of search results by prioritizing the links which users would find most helpful.

Developing wisdom is an iterative process and comes with a lot of dialogue. It takes time. At an individual level, it may be an internal dialogue. If done at a team level, this will be a team dialogue—to hear and mesh all the perspectives, to develop insights and then align to achieve goals. When a few people in a team interact, they share their thoughts, which may be divergent. Listening to another person with respect, engaging in a healthy challenge and then

converging leads to superior decision-making, which creates significant impact. To me, this process and rise in team knowledge via dialogue is team wisdom.

Diversity enables wisdom, as it simulates learning from the experience of others. This explains why progressive organizations focus on building diverse teams—with representation of gender, ethnicity and maybe age as well. Bringing together people from various perspectives helps develop the wisdom of teams. At an individual level, seek out and develop relationships with those who can act as a sounding board. I personally find a sounding board to be invaluable. And of course, my biggest sounding board is my wife, and I surely get a lot of diverse and challenging views there!

When this idea of collective and collaborative wisdom extends beyond individuals and teams in an enterprise, it becomes organizational wisdom. This helps create an environment that nurtures trust, promotes open communication and values diverse viewpoints, becoming an integral part of the organizational culture.

The organizations who focus on building this organizational wisdom have their leaders exhibit qualities such as self-reflection, serving as role models, showing compassion, taking responsibility and being accountable for their actions. These organizations understand that wisdom is not just about what they achieve, but how they achieve it and the positive impact they have on society as a whole.

Wisdom has been a timeless and intrinsic human capability and is now perhaps even more relevant in the data-first world and AI age in the twenty-first century. It takes years to build wisdom through our experiences and knowledge. It serves as a guiding light amid the overwhelming flood of information, helping us discern what truly matters and make effective decisions and is likely the key to the future of our interconnected global society.

Key takeaways

- In the Big Data world, wisdom's significance has not diminished; in fact, it has grown even more vital. It acts as proprietary knowledge, setting individuals apart and enabling them to make intelligent and more effective decisions.
- Defining wisdom is complex due to its multifaceted nature. It comprises three core elements: thinking deeply, upholding strong values and making sound, well-informed decisions that have a lasting and sustainable impact.
- The basic ingredients of gut instinct and wisdom are the same, but wisdom is more reflective while gut instincts are more spontaneous. Gut instincts can become highly reliable when they move from unconscious incompetence to conscious competence.
- Building wisdom is anchored in five key interconnected principles—humility, seeking diverse experiences, recognizing patterns, reflection, teaching and sharing. They influence and enrich each other to create a virtuous cycle of cultivating wisdom. Reflection and self-awareness can be nurtured through practices like meditation.
- Team wisdom surpasses individual wisdom by combining unique perspectives, forming collective wisdom. In organizations, this evolves into organizational wisdom, fostering trust, open communication and diverse viewpoints, ultimately shaping the organization's culture.

Data Sharing vs Data Privacy

Find the Right Balance

'It is not the data that is being exploited, it is the people who are being exploited.'

—Edward Snowden,

Former CIA employee and National Security Agency contractor

We saw in the previous section how data security is one of the biggest challenges for enterprises in the data-first world. We also saw that there is a trade-off and balance between data democratization and data security for enterprises. In a similar way, data privacy is a big concern for individuals and there is a trade-off between data sharing and data privacy.

The topic of data privacy has been a subject of intense debate since the inception of the Internet. While data serves as the basis to deliver highly personalized services, as discussed in Chapter 21, The World of Hyper-Personalization, it also raises significant privacy concerns. It can be incredibly beneficial in many scenarios but equally disconcerting in others. It is quite unnerving to realize that there is someone (even if it is a server located somewhere) who knows who you are, what you like to do and how you like to behave. The issues are not easy to resolve because of the complex interplay between individuals, organizations and government owing to diverse interests and objectives, which are often not aligned with each other.

While a unified global consensus on the matter remains elusive, it often comes down to individual choice and responsibility regarding how much data one is willing to share in the digital age. People always say that if it is private, it's best kept off public platforms like

do not put it on Facebook. This advice echoes age-old wisdom that cautions against divulging personal information, even to those closest to us. However, in the hyper-connected world we live in, it's becoming increasingly challenging to discern when, how and what data is being collected and, crucially, how it's being used.

The privacy issues become even more complicated when you are dealing not just with an organization but the government. We rely on our governments to make many decisions for us, but sometimes, those decisions might come with inherent privacy concerns. The following example from China highlights the two sides of the data sharing and privacy debate—the benefits and the concerns. In China, the Ministry of Public Security, China's top police agency, introduced a social credit system to enhance trust in Chinese society. The system began with a focus on financial creditworthiness to help individuals and businesses make fully informed decisions. But in recent years, the system has moved beyond financial credit worthiness to judging citizens' behaviour and trustworthiness. Social credit rankings became the basis for rewarding or punishing a citizen. For example, if you don't pay a court bill or play music too loud, you may lose certain rights like using a subway or booking a train ticket and much more. The surveillance state in China continues to expand as authorities use phone trackers, DNA samples, iris scan samples and voice prints to build comprehensive profiles on citizens with the aim to apprehend suspected criminals. Nowhere in the world, such extensive data systems at this scale are used to control individual behaviours.¹

In a less intrusive but extensive effort, the Aadhaar system was implemented to streamline government services and improve access for Indian citizens. Aadhar provides citizens with a unique identification number linked to various services, including taxation through PAN cards and banking services offered by private entities. In the United States, the Social Security card has historically been used for identification and to provide access to government benefits.²

While such government initiatives do ease your life or bring transparency, they have also led to personal information being in the hands of organizations and governments, which has also sparked debates about privacy and security. They also put your information outside your personal control and susceptible to hacking, which can have an adverse impact on individuals, for no fault of theirs.

Such examples highlight the importance of maintaining a delicate balance between enhancing services and safeguarding individuals' personal data in an increasingly digital world. Clearly, these are not easy choices to be made and there are multiple stakeholders involved, making it even more difficult to achieve.

The complex dynamics of data privacy: Individuals, organizations and government

Most times individuals are oblivious to the gravity of privacy issues. One key reason for this is that many individuals are not adequately informed about the risks and implications of data misuse. Often, individuals do not fully understand the value of their personal data or the extent to which it is collected, shared and potentially exploited by various entities.

Most individuals are often unaware of their data privacy rights, which include the ability to control their personal information and understand how it's used. Without proper awareness, people may unknowingly agree to data collection and sharing practices they would object to if they were better informed. Additionally, individuals may not take proactive measures to safeguard their data because they underestimate the risks associated with data breaches or misuse. It's common for individuals to recognize the importance of data privacy only after experiencing incidents like data breaches or identity theft. By that time, significant damage might have already happened.

This leads to information asymmetry, when one party in a transaction possesses more or superior information than the other, which many organizations have been benefiting from. A perfect

example is Mark Zuckerberg's congressional testimony regarding the Cambridge Analytica breach in April 2018. Senator Durbin asked, 'Mr Zuckerberg, would you mind telling us which hotel you stayed in last night?'³

To this, Zuckerberg responded, 'No, I am uncomfortable sharing it, Senator.'

Senator Durbin continued, asking, 'Would you like to tell us whom you messaged last night?'

Again, Zuckerberg replied, 'No.'

Senator Durbin then pointed out, 'That is what this (privacy issue) is all about ... what information Facebook is collecting, who they are sending it to, and whether they asked my permission to do that?'

In response, Zuckerberg explained, '... because ... people choose.'

It is hard for me to get my head around this because we all know, the user who gave consent, did not know what s/he was consenting to and did not really CHOOSE.

We all know of cases of organizations being fined heavily for lack of transparency in the use of individuals' data. Google was fined ~\$57 million in 2020 by the French data-protection authority for failing to acknowledge how it used users' personal data.⁴ In 2019, X (formerly Twitter) admitted to letting advertisers access its users' personal data to improve the targeting of marketing campaigns.⁵ While these instances create a stir in the world, they rarely result in organizations changing their ways for the longer term. We all take these revelations about companies with a pinch of salt. We know it is happening and we have an option to stop using the apps or the products being offered, but how many of us really do?

Individuals have also been a part of this ecosystem of data-misuse. In 2015, a Morgan Stanley financial adviser pleaded guilty to saving the data of ~730,000 accounts, attempting to take data with him to a competitor. External hackers accessed the employee's PC and briefly posted personal data of ~900 accounts online.⁶ But as individuals gain more understanding of the risks and consequences of such actions, these kind of data breached can be avoided.

It's even worse when the government starts using data to target individuals. In 2014, the Chicago Police sent uniformed officers to make 'custom notification' visits to individuals they had identified as likely to commit a crime through a computer-generated list.⁷ In the US, the Restrict Act aims to ban TikTok, but it also grants the government unrestricted access to a wide range of our data, including our devices, security cameras, internet history, payment apps and more.⁸ Such instances and the China story I mentioned above, feel right out of a Hollywood movie—like Tom Cruise in the Hollywood movie *Minority Report*, where predictive policing was the anchor of the plot.⁹ A cautionary tale about the potential erosion of privacy and civil liberties, where the police department uses precognitive abilities to predict crimes, an officer was forced to become a fugitive when he is accused of a crime he hasn't yet committed.

The implications are deeply troubling as it undermines the Freedom of Information Act and permits unchecked government access to sensitive personal information, including medical records, financial data and private conversations.

During my time working in policy, particularly during my close collaboration with NASSCOM, the Indian Association for Software Companies, from 2011–16, I gained valuable insights into the complex dilemmas that policymakers often face. One of the key challenges is the balancing act between optimizing policies for the greater good versus the interests of individuals. This raises critical questions around data ownership: should it belong to the organizations whose platforms the data is shared on, or should individuals have ownership rights over their data?

The policymaker who is looking at the greater good of society and chooses a path that helps most may find themselves at a crossroads. Let me illustrate with an example. In the US, policymakers made a choice to make clinical data freely available to researchers, aiming to advance society's medical knowledge. While this decision is commendable, it raises ethical questions. Some individuals may not have wanted their sensitive data to be exposed. If they objected,

would they be excluded from clinical trials? On the other hand, companies invested heavily in these trials and might have preferred to keep the data for a competitive edge. But does that make it ethical to use that data that belongs to an individual in the first place?

In this scenario, what is the role of policymakers? The dynamics at play involve a triad of government, organizations and individuals, each with their own set of incentives and priorities. These interests rarely converge, making it difficult to establish standardized policies or a clear direction in the realm of data privacy.

While the policymakers are on the fence on what should be done or can be done, people are often left to navigate the complexities of data privacy on their own, having to make individual choices about how to protect their personal information. This complex interplay highlights the challenge of formulating unified data privacy regulations and achieving a consensus on the responsible and ethical use of data.

Call for global standards for data privacy

To me, the biggest question in this debate is who has the right to the data—is it the person whose data it is or the organization who created a method and made the effort to generate the data. Or it's the government that is working towards making things for greater good. Once we clarify the data ownership, the contours of its use and misuse will become clear.

Earlier in Chapter 17, *Organizational Alignment*, we have spoken at length about data ownership—from an organization's perspective. From an individual perspective, this is possibly the hardest question to answer, and probably that is why, there is not a single uniform code around the world on data privacy. Each industry, state/central government, has taken their own approach on this, and it also varies by the kind of data; for example, HIPAA (Health Insurance Portability and Accountability Act) for privacy of medical data in the US, GDPR (General Data Protection Regulation) for regulation in the European Union (EU) on data protection and privacy, CCPA (California

Consumer Privacy Act) for residents of California, United States. In 1981, Council of Europe's Convention for the Protection of Individuals tried to bring in a uniform data-security framework for personal data through a treaty which was signed by fifty-five countries.¹⁰ But for various reasons like lack of reinforcement mechanisms, disagreement in approach across countries, national interests and more, countries have not uniformly followed through on their commitments, rendering the treaty largely ineffective.

I believe there is dire need for a global standard for privacy. All of us have experienced blindly clicking 'accept' to pages of information that show up when we are trying to sign up to certain platforms—without reading. Social media giants change their privacy settings every few months, and we accept. There is a trust in the system—that someone out there is looking into my privacy, my safety. The truth is, there is no one. There are no global standard privacy laws currently and this has a possibility to impact many across the world adversely. We need global standards of privacy that are simple to understand, and the user knows what they are signing.

Although GDPR is emerging as a front-runner in data-privacy standards with the EU (twenty-eight countries), North America and Japan being GDPR-compliant, it is still not a global standard. I see value in looking at privacy laws in countries like Iceland—called the 'Switzerland of data'—which took massive steps in creating strict data privacy laws: under its 2000 Data Protection Act, amended in 2010 after the financial crisis, data can only be collected for specific purposes with the subject's informed and unambiguous consent. They must be informed about the data type, collection purpose, processing method, data protection and the ability to withdraw consent at any time.¹¹

Having said that, most data-privacy laws today are complex and often hard to understand. They are technical, and each organization interprets them as they want. The laws continue to be murky and hard to understand. They are broad and maybe hard to implement.

I believe, since these laws impact the mass population, they need to be simple, and need to protect citizens. Since 1979, Global Privacy

Assembly—an international forum that brings together data protection and privacy authorities from around the world to discuss and collaborate on issues related to data protection, privacy and information rights (now 130 members strong)—has been advocating standardization internationally in data protection and privacy laws.^{[12](#)} However, these standardization efforts are still happening in silos and within a group of experts. A non-technical person (and even technical people) finds it hard to read the fine print.

In my view, data-protection conversations need to be elevated to higher levels of global governance. Leaders of the world have to come together to gain alignment around this topic. Just like the war for boundaries and the war for arms, data war is an emerging war—a turf that each nation will want to protect for its citizens. And it will become increasingly harder in this globalized world, where the internet has connected everyone. Early signs of this are visible, with countries banning apps like TikTok which were collecting data (of its users) across the world. Imagine the data owned by Facebook, Google and the likes. The number of data breach lawsuits against Facebook around the world is rising, suggesting the rising discomfort within various government authorities.

Data is already being used to impact key decisions of nations as seen in the Cambridge Analytica scandal, where Facebook data of 50 million users was used to profile voters.^{[13](#)} Voting decisions were swayed with hyper-personalized messaging. The biggest economy of the world may not have had fair elections, as data was used to influence decision-making. Such is the power of data and the potential significant impact that absence of data privacy laws can result in.

I anticipate, as awareness of the extent of control that can be exercised with data becomes visible, data wars will form a part of modern warfare. Data, when used at a mass scale, has an ability to manipulate the decision-making. Elevating these conversations at the highest levels of governance and coming to a common ground is important.

An individual's choices around data sharing and data privacy

As is evident, establishing global standards, while critical, is a complex and difficult task to achieve and thus remains work-in-progress, given the complexities involved and varying national interests. Meanwhile, we continue to live in a world where an individual's data is being used for varying purposes. Although there is enough and more in the world to debate on this topic, as I highlighted above, the individual has limited options. If the person is to enjoy the benefits of the data-abundant world, and wants to be a part of it, the person will need to make their data available. Can one stop how this data is used? Not necessarily. There is limited opportunity.

I am sure you all have also experienced this. Each time you go onto a website, you are faced with pages of terms and conditions, which you need to mandatorily sign if you wish to access some content. How often have you felt helpless and incompetent to read those pages? A 2019 survey estimates that 78 per cent of Americans agree to the terms and conditions without even reading.¹⁴ Those who read, do not have an option to question or challenge the terms or access content without signing. Giving access to your data is the price you pay for being able to participate in this data abundant world.

My personal favourite is this example of Truecaller. Truecaller is an app that you install on your phone so when someone calls you, you are able to see their name, even if they are not stored in your phonebook. Incredible, isn't it? Do you wonder how they do it? When you download their app, all your information, such as your name, phone book contacts, operating system and device ID, is promptly transmitted to Truecaller's server. It gets displayed on your screen when someone calls you or you look the number up using Truecaller search. But as stated in their policy, they also collect additional information from your phone, including your contact list, your geo-location, IP address, device ID or unique identifier, device

manufacturer and type, device and hardware settings, SIM card usage, applications installed on your device, ID for advertising, web browser and a lot of other information. The full extent of what they do with your data is unknown.^{[15](#)}

Many of you may not even know that so much of your data is now becoming the property of an independent private entity, when you download the app. While the company has access to your data, you have no transparency on how it is being used and how safe it is against breaches and thefts. In May 2019, it was alleged that data of over 300 million Truecaller users in India (which makes up at least 60 per cent of Truecaller's total customer base globally) was being sold.^{[16](#)} Although the security breach was denied, it raises a critical concern around safety and security of our personal data.

I am not advocating that you remove Truecaller, I still use it and find it helpful but be aware that the choice of data sharing and data privacy is yours.

So what can an individual really do?

While it may be difficult for individuals to monitor every aspect of data access, vigilance is essential. Identify the critical data points that are most sacred to you and could be easily misused. Some key principles to keep in mind:

- **Avoid oversharing:** Exercise extreme caution when sharing critical information online, only providing necessary personal information identifiers (PII) for transactions.
- **Share data with trusted brands and platforms:** Reduce the risk of your personal data being misused by sharing it only with trusted entities.
- **Anything posted online, stays there forever:** Anything shared on social media becomes public property, and you have no control whatsoever on who

has access to it and how it can potentially be used with malicious intent.

- **Beware of banking and financial fraud:** Never share bank details unless you're certain about the purpose and use secure browsing for internet banking. Never share sensitive information like passwords, credit card pins, etc.

Ethics around data usage and data ownership: My view

My perspective on this matter comes from a practical acceptance of the reality of the digital age: data sharing is unavoidable and essential in the data-first world. So instead of avoiding it, individuals must learn how to do it intelligently and prudently.

Controlling every aspect of data usage might not be entirely feasible, but you must know where to draw the line. Making informed choices about what to share and what is sacred to you is paramount.

So at an individual level, it's crucial to be mindful when sharing information because once data is out there, it remains accessible indefinitely. The key here is to discern what is truly private and avoid sharing it on public platforms like social media. Be judicious about sharing critical information while benefiting from the opportunities presented by the abundance of data.

A belief I hold firmly is that the one who controls information owns it. Control, in this context, doesn't solely pertain to data collection but also encompasses access and usage. Governments worldwide are increasingly involved in data governance, emphasizing their role in controlling and regulating access to citizen data.

For your own data:

- Data stored on your personal device/server: You own it.
- Data shared (with consent) with hardware/software providers: The provider owns it.

- Data uploaded to the internet and made publicly available: Your data is now public and accessible to anyone.

Striking the right balance between data sharing and data privacy is a complex challenge. The decision you face is whether to embrace the opportunities of the data-driven world, which may require sharing data, while being aware that the data you share is no longer solely owned by you. The benefits of this data-driven world come with the trade-off of reduced personal data control, and it's a paradox we encounter daily. Ultimately, the choice rests with you.

Key takeaways

- In the digital age that we are in, we are sharing data like never before. The obvious paradox that emerges as a consequence is that while data sharing is inevitable and essential, it also gives rise to significant privacy concerns and challenges.
- The data privacy issues are complex as they involve interactions and conflicts between individuals, organizations and governments, each driven by their interests and objectives, which are often not aligned with each other. To solve these divergences of interests requires unified regulations that balance the interests of all parties.
- There is a lot of complexity and divergence in data-privacy regulations, which makes them difficult for an individual to navigate. Therefore, there is a pressing need for global privacy standards, which requires leaders worldwide to elevate discussions to the highest levels of governance for achieving alignment on this critical issue.
- Establishing global standards is still an ongoing process. Meanwhile individuals have to make their own choices to best control how their data is being used. Therefore,

individuals should exercise caution when sharing personal data, avoid oversharing and limit data sharing to trusted brands and platforms.

Digital Engagement vs Mental Health

Connect with Your Inner Self

'Technology improves the lives of people who can avoid being dominated by it and forced into debilitating addiction to it.'

—Frank Kauffman,
Scholar, educator, innovator and activist

The digital world has presented numerous opportunities for individuals, offering the potential for a better life in various ways. Engaging with the digital world has enabled us to connect, learn and work in ways that were once unimaginable. However, much like any facet of life, it comes with both positive and negative aspects. On the one hand, engaging with digital brings forth countless benefits, but on the other, it overwhelms us with constant stimuli and information, adding to the complexity and fast-paced nature of modern existence.

This highlights another paradox of the digital age. The benefits of digital engagement are undeniable. Undoubtedly, technology has enriched our lives, offering increased convenience and efficiency. However, the same digital engagement, if excessive and constant, can lead to burnout, stress and a sense of being overwhelmed, necessitating frequent disconnection and detox from it. Because if we permit technology to dominate our lives, we risk compromising our mental well-being.

In a UK-based research, around half of the respondents (45 per cent) said that the stress of data overload affected either their sleep or relationships with family or colleagues, and one in three respondents said that it made them feel anxious, fidgety and unable

to relax.¹ How vast and varied are the impacts of digital over-engagement. Therefore, there is a growing need for moments of disconnection and digital detox to restore well-being and mental clarity.

In this chapter, I will play the devil's advocate, shedding light on the negative impact of digital in our lives. Because I wouldn't be doing justice to the topic of this book without acknowledging that there is a flip side to the digital world we are in. And while I highlight some key issues, I will provide you with some practical advice and age-old wisdom on how to deal with them effectively, so that every individual can fully reap the benefits of the data-first world and the AI age.

Rising digital engagement, designed to keep you hooked

Brands and organizations are making use of digital engagement and reaching out to individuals in multiple ways. The trend of hyper-personalization we spoke about in Chapter 21, The World of Hyper-Personalization, reflects on the increased efforts by brands and individuals alike on raising digital engagement. On an average, a typical internet user is spending 40 per cent of their waking hours—that is, almost seven hours per day—using the internet across devices in 2021 as compared to six hours in 2013.² The average screen use in the US, especially amongst kids and tweens, has increased by 17 per cent between 2020 and 2022.³ Social media has become an integral part of how we engage with the world. In 2021, approximately 56 per cent of the global population were social media users. This share is projected to increase to 74 per cent by 2026.⁴ Entertainment has gone digital which is evident in the fact that in the OTT video segment, the number of users is expected to reach 4.2 billion by 2027.⁵ Online shopping is the new way of shopping: In 2025, e-commerce sales are estimated to account for a quarter (24

per cent) of the total global retail sales, up from 19 per cent in 2022.⁶

As individual lives are significantly transitioning from offline to the digital world, brands are actively competing for a slice of their screen time. This is done by purposefully engineering the engagement to capture and retain an individual's interest. User-friendly design, personalized content, notifications and gamification are all aimed to keep users actively involved and willing to spend more time online. Brands get the opportunity to use these platforms to convey their messages, sell their products and services and monetize through user interactions. Additionally, capturing more and more customer data is critical for organizations to deliver highly engaging experiences. But this would only happen when people are hooked to the portals and stay that way. Hence, the content is made to deliver instant gratification, appealing to our senses.

Rising complexity and ill-effects of excessive digital engagement

The complexity of the world we live in today is on the rise. At the time of starting my career, I would get emails and knew that was the only official channel for communication. Now, I get emails, WhatsApp, phone calls, LinkedIn messages and more, which often results in me not having enough time to respond to each one of them. I am always on high alert, trying to be on top of all my messages across multiple channels.

Drew Barrymore's character in the movie *He's Just Not That Into You*, laments, '... it is confusing. I met the person on a dating app, we scheduled our first date via SMS, he left me a voicemail for a plan and broke up with me on an email. I really do not know. Life used to be simple, now I have to keep track of all the possible places he must be sending me a message to be able to have a dating life. I need to check my SMS, my WhatsApp, my email and my voice mail. Earlier, people would just show up and meet. Love used to be simple then.'

Of course, there are so many benefits. For example, it has become extremely easy to plan your travel online, but at times smartphone apps and multiple websites can easily lead to information overload making the whole process overwhelming too. Back in the day, planning a vacation was straightforward. You'd book a flight, reserve a hotel mostly through an agent and maybe buy a guidebook. Now, it's overwhelming as you get bombarded with information from multiple sources! You research destinations on Pinterest, search for accommodation on Airbnb, chat with the hosts on WhatsApp, receive travel itineraries via email, get restaurant recommendations on Yelp and check in with your friends on Instagram to see their travel photos. It's like I'm coordinating my trip across a dozen apps and platforms. Sometimes I feel exhausted even before the vacation starts!

Interestingly, there are apps that are solving this problem too, like the aggregator apps that aim to provide a one-stop solution for all my travel needs—but the problem is real. The complexity in our lives is on the rise as a result of these multiple channels of digital engagement. And as these complexities grow larger, there are several adverse effects that are emerging, significantly impacting the quality of our lives.

First and foremost are the physical health issues that are increasingly becoming commonplace. Prolonged screen time can lead to physical problems like eye strain, poor posture. Excessive screen time can be linked to a decrease in sleep duration, insomnia, poor sleep quality and daytime issues like drowsiness and irritability. Excessive screen time can also take away time that could be used for physical activity, which negatively impacts the health of an individual in the long run.

Another prominent ill-effect of the excessive digital engagement is the increasing disassociation with the physical world. Spending more time communicating via digital platforms and social media often leads to a decline in the real-world interactions. Lack of face-to-face social interactions diminishes an individual's ability to engage in meaningful, in-person relationships. A preoccupation with digital

engagement can divert attention and time away from real-world activities, such as hobbies, sports and outdoor experiences.

Added to this multi-channel engagement are new forms of evils—like cyberbullying, cyber-stalking. This rising complexity, if not managed, will start impacting our mental health. It creates more stress for things that may not be necessary.

I am sure, you all would have at some point sent a message on a digital channel like WhatsApp, observed a 'seen tick' and then stressed about the fact that the receiver is not responding. We do not talk about this often, but we all feel it. Some people are affected by it, others are not. But it is information that has a possibility to trigger people.

Many of us have, at some point, measured our professional and personal lives by the yardstick of our social media popularity. A popular mantra of the social media networking age is, 'If it isn't on social media, it hasn't really happened.' And if it's not liked by many, it isn't worthwhile. It is one of the many ways in which excessive digital engagement can lead to mental health issues like anxiety, depression, feeling of inadequacy and reduced attention span.

As all these digital platforms are designed to make them easy to use, high engagement happens naturally. This high engagement on digital channels and digital media, which is often filled with images and stories that are aspirational, and often not achievable, leads to depression, disturbed sleep patterns and distorted body image. Some of the common negative psychological effects of digital overuse are depression and anxiety, isolation, FOMO, lack of self-worth.

Digital overuse may impact mental health, especially leading to an attention deficit disorder (ADD). Instant gratification, dopamine rushes due to what you see, and then switching between content, is leading to lower attention spans. This is significant, as I state in Chapter 23, Information and Wisdom, that wisdom is dying, and one of the main reasons is that we are not giving enough time for reflection. Reflection needs focus and staying with a problem. With ADD on a rise, wisdom is bound to be on a downfall. A 2018 longitudinal cohort survey study of adolescents aged fifteen and

sixteen years revealed that frequent digital media use may increase the risk of having symptoms of ADD by about 10 per cent.⁷

I do paint a dark picture, don't I? While most of us hover somewhere on the less extreme side of the mental health spectrum, stress being a common side-effect, it is undeniable that the extent of damage is intensifying with each passing generation, in fact with each passing day. Thus, it is high time we are aware of it, acknowledge that the problem exists and proactively do something about it. Easier said than done, I agree. But I believe that with discipline, effort and a bit of wisdom, it can be achieved.

Simplifying in the data-abundant world: Digital detox

Over engagement is also a form of addiction. Just like any other form of addiction, the individual doesn't realize how and when an activity that they partake in for leisure or entertainment becomes a compulsion. So much so that it starts impacting their mental well-being. And many of us, unknowingly or due to the nature of our modern living, are getting addicted. So it's essential to proactively undergo a conscious detox to maintain a healthy balance.

Five steps to moderate digital engagement

Counter Technology with Data

- Use technology to monitor usage
- Analyse the usage pattern
- Prioritize area of improvement

Create Boundaries

- Take out time to voluntarily refrain from using digital devices e.g. digital detox programs

Engage with Physical World

- Taking up outdoor physical activities
- Cultivate a hobby
- Opt for phygital experiences



Seek Help

- In case you are not able to deal with it on your own, seek help from friends and family or professionals

Connect with Inner Self

- Connect with your inner self and disconnect to unleash the positive energy and enhance productivity

While a lot has been written about the negative impact of the digital world, a lot of control is in the hands of the individual. Every individual can manage how to engage with the digital world. How much data a person engages with so that it does not become an addiction, is dependent on individual discipline. Granted, it is tough, and one does not realize when the data consumption starts rising, and that's why maintaining discipline is so crucial. Some practical ways that I have observed, that help simplify and manage how you consume and engage with data are:

Counter technology with data

The first step to achieving any goal is to evaluate where you are. Here data can empower you to make informed decisions. By accessing and analysing data using the technology at your disposal, you can MAP—monitor, analyse and prioritize—your digital engagement patterns and find ways to tackle it:

Monitoring for self-assessment: Begin by assessing your current digital habits and usage patterns. Ask yourself how much time you spend on screens and whether it aligns with your health and well-being goals. There are many applications available today designed to track screen time and social media usage. These apps often provide insights into your digital behaviour. Extend the monitoring to other vital activities, such as sleep, exercise and meditation. Use dedicated apps or wearables to track and record your progress in these areas.

Analyse your digital footprint: Most times we are unaware that we are spending too much time on screens. How often have you begun scrolling through digital content, only to realize that you've spent an hour or more engrossed in it? So, analysing the data collected by monitoring apps can help you gain a comprehensive view of your digital footprint.

Prioritize areas of improvement: The comprehensive view of your digital engagements can help you identify areas where you may

be spending excessive time and determine which of these areas have the most significant impact on your overall well-being. This recognition allows you to focus your efforts on making constructive changes in the areas that matter most.

Once you have MAP-ed your digital engagements effectively, you take some practical steps to bring positive changes. At times, just taking purposeful short breaks from the phone or computer helps. Some practical ways I can imagine this can work out is:

- **Practice digital abstinence:** Challenge yourself to a brief period of complete digital abstinence, ranging from a day to a week. Or pick one day of the week as a device-free day on an ongoing basis.
- **Specific detox:** If a particular app, site, game or digital tool consumes too much of your time, implement restrictions on their usage.
- **Social media detox:** Focus on reducing or eliminating your social media presence for a specific period of time.

Create boundaries

Sometimes, digital over-engagement can also be a matter of necessity. For example, a person working as an IT expert, or a software developer, has to work on screen for a large part of their day. And there is no escaping it. In such scenarios, it is absolutely essential to create personal rules and boundaries for screen use and adopt a structured approach to balance digital engagement.

I have realized that defining some boundaries for yourself, communicating, and sticking to those, is really important. For me, personally, a structured digital detox programme works wonders. I have been going for Art of Living^{*} silence retreats at least once a year for more than twenty years now. Each time I am completely disconnected from my work for five days. That is my sacred 'me' time, marked for contemplation and self-reflection. That is the time when I set boundaries and cut off the external noise. Will my

business stop running? No. But taking that break always makes me come back recharged and revitalized.

I have friends who have chosen not to be on the smartphone bandwagon and stick to feature phones. Yes, they will not get the 100 per cent benefits of what the digital and the data world has to offer, but they prefer to have control over when to be a part of it and when to walk away from it.

Engage more with the physical world

Staying connected to the real world is essential for the mental well-being of an individual. Balancing physical activity with digital engagement is essential for a well-rounded and fulfilling life. There are multiple ways to increase your interactions with the physical world:

Taking up outdoor activities: Outdoor activities like nature walks, hiking, traveling to new places or simple activities like gardening will help you engage with nature and natural surroundings, significantly reducing time in front of a screen. I have a beautiful garden at my house, personally cultivated and maintained, which I work in every morning before my busy work schedule takes over. It not only helps me connect with nature, but it gives me a real sense of pride and satisfaction to see my favourite plants grow and bloom.

Taking up physical activities: Whether it is exercising of any kind, or sports, workout or dance, incorporating physical activity can be a real motivator to temporarily disconnect from the digital world. My son, a digital native, was once glued to the computer, excelling in video games. Fearing digital addiction, I encouraged him to engage in real sports. In recent years, he started pursuing squash seriously, and his computer time was substituted with intensive training on the court, replacing his sedentary lifestyle for an active one. Not only is he developing skills and getting better at the sport, but he is also getting fitter, stronger and happier, and the digital use has gone down dramatically.

Cultivate a hobby: An effective stress buster for an individual is to engage his creative faculties. Pursuing hobbies and interests that involve hands-on activities, such as painting, cooking or playing a musical instrument can help individuals develop their creative instincts and provide them with avenues for creative expression. It is an effective way to detox from the passive interactions from the digital world.

While over-engagement with digital is a problem, merging the physical and digital worlds to derive a 'phygital' experience can be the solution. By leveraging digital apps that complement physical activities you can enhance your experience, track your progress more effectively, set and achieve goals effectively.

Of course, there can be many other activities an individual can engage in, but the essence of it all is to find ways to disconnect and detox to avoid digital over-engagement.

Seek help, if needed

When over-engagement reaches the point of addiction, sometimes it takes another person to help you work through a problem. As per a Pew Research Center Report, 46 per cent of Americans said they could not live without their phones, indicating serious addiction requiring professional help.⁸

Seeking help, professional or otherwise, is sometimes the answer. Experts often are able to pinpoint things that we are unable to see, but experience. They are able to identify which of our experiences are not in line with the norm. To have someone, who has seen situations like yours many times and can guide you. Counselling helps, as it empowers one to reconnect with people one has been unknowingly distancing from. To be a part of a supportive community that understands what one is going through. Again, there are various ways you can seek help:

Joining support groups or communities with individuals facing similar issues can also offer shared experiences and coping mechanisms. Family and friends can provide emotional support and

encouragement to reduce digital engagement. Self-help resources like books, articles and online resources dedicated to digital addiction recovery can be valuable self-help tools. Some individuals may benefit from life coaches who specialize in digital well-being and productivity.

Sometimes, when your mind becomes your enemy, seeking external help to control the mind might not just be helpful but necessary.

Connect with your inner self

We are in the age of hyper-consumerism, where our relentless pursuit of external symbols of success such as wealth, possessions, power and status has trapped us in a never-ending cycle of desires and expectations. This, combined with the overwhelming stimuli of the digital age, has dire consequences. The stress levels are increasing, tolerance and empathy is decreasing, relationships are suffering and the moral fabric and integrity are weakening.

The digital age has introduced overwhelming complexities with constant external stimuli and information overload. There is a lot of noise, too much chatter, too much information to assimilate and respond to. Amid all this noise, when one starts looking inside, the noise quietens down. You are able to discriminate the signals from noise. There is a calmness. The constant stimulus of the external world brings complexity and makes us frenetic. While the calmness that comes from going deep within self brings simplicity of thought, decisions and relationships. One is able to anchor oneself in the internal self and cut the external noise.

But how do we achieve such connectedness with oneself? How do we really go deep inside and connect with our inner self, to be able to anchor and simplify our life?

Spirituality holds the solution

Let me be clear, I am not asking you to take extreme vows and renounce your worldly possessions and live like a hermit somewhere

deep in the forests or high up in the mountains! I believe spirituality, a concept that is often misunderstood, if practiced regularly, can be an effective tool to bring inner calmness, clarity and a sense of purpose into our lives.

Spirituality can mean different things to different people. For some, it can primarily be about believing in God and being involved in organized religious practices. For my father, chanting mantras from the holy book gives him a sense of anchor, and he often feels agitated on the days he is unable to do so. He says it helps him find peace and hope in the greatly complex and uncertain world around us. For some, it can be about non-religious experiences that enable them to connect with their inner self through solitary moments of reflection, or immersing themselves in the tranquillity of nature, or embracing practices like yoga or meditation. It can also be through regular practice of your favourite sport or activity, where the sheer act of engaging in something you're passionate about provides a heightened sense of self and moments of clarity. The common thread here is the pursuit of a deep and calming activity, a process of going inside that enables self-reflection and a more profound understanding of oneself and the world, regardless of the specific path chosen.

My favourite way of connecting with my inner self is through meditation, which I have been practicing for over thirty-five years. But I am equally passionate about sports and feel the same calmness and centredness after an intense game of squash, as I do after meditation!

Connecting to your inner self is not just about sitting on a couch and closing your eyes, every day with discipline, but practicing whatever you like to do with utmost mindfulness, improving at each step like sportsperson do. I recommend you find your own path. A path where you feel one with yourself and feel connected to yourself at a deeper level. Your method may be personal to you, but the goal is universal, to be connected to oneself at a deeper level.

In essence, to find true peace and happiness in today's fast-paced material world, it is imperative that we discover spirituality and connect with our inner self. There is a great power that lies deep

within all of us. This power is also a universal force that also connects all of us. We look for solutions for our problems outside, whereas all solutions lie within us.⁹ The answer to a very twenty-first century problem of a hyper-digital world might well lie in the timeless wisdom of our ancestors!

Small changes to build habits

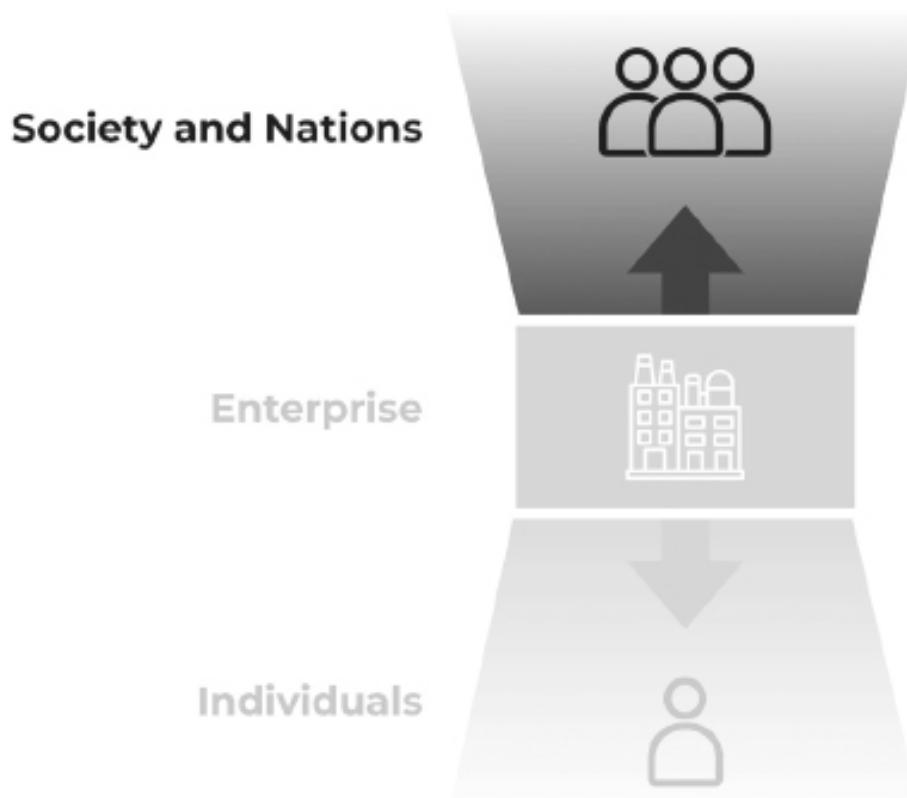
Like any substantial impact, achieving a balanced digital engagement is a process. It will not happen instantly, or in a few days. To bring significant and sustainable change, you have to make it a way of life. Which means, anything attempted, should be done continuously, and for a long time to see tangible results. After all, 'Rome wasn't built in a day'.

Just like the two-speed approach I suggest in Chapter 14, *Agility*, that an organization needs to build long-term capabilities while achieving short-term goals, for an individual as well, the key is to start by making **small changes**. Take on easy and achievable targets that you believe will help you achieve some instant benefits and convert them into a **habit** by gradually adding to it to achieve the long-term vision of a healthier, well-rounded life. Because if you take on big targets, you may end up doing nothing about them. Just like the New Year resolutions we take every year. For instance, if you've been struggling with anxiety and stress, attempting to meditate for an hour may become overwhelming and unattainable. A more effective way would be to begin with small mindfulness practices like dedicating just five minutes a day to deep breathing exercises or short meditation sessions, which can have a calming effect. As you progress, you can gradually increase both the duration and intensity of your practice. Then over time, you can explore additional coping mechanisms, such as engaging in a physical activity on a regular basis or enrolling in structured programmes to make these practices part of your routine.

Key takeaways

- One of the paradoxes of the digital age is that while digital engagement brings multiple benefits, excessive and constant digital engagement can lead to burnout, stress and a sense of being overwhelmed, necessitating digital disconnect and detox.
- To counter the negative effects of excessive digital engagement, start by monitoring your usage, use data to analyse your digital habits and prioritize areas for improvement. Practical steps like digital abstinence and reducing social media usage can help achieve a healthier balance. In many situations, adopting structured approaches, like digital detox programmes might be needed.
- It's crucial to engage more with the physical world by participating in outdoor and physical activities, spots and cultivating hobbies, thereby enhancing overall well-being and reducing excessive screen time.
- When digital engagement turns into addiction, seeking professional help, joining support groups, involving friends and family, or utilizing self-help resources can be effective strategies to regain control over one's digital habits and reconnect with the real world.
- Reconnecting with your inner self through spirituality, especially meditation, the timeless wisdom of our ancestors, can bring inner calmness, clarity and simplicity to navigate the external chaos, fostering personal growth and well-being.

SOCIETY AND NATIONS



'Data transforming the world'

Data Collaboration for a Better World

A New Vision for Global Collaboration

'Data sharing is a powerful tool that can be used to solve some of the world's biggest problems.'

—Bill Gates,
Co-founder and former CEO, Microsoft

In this section so far, we have discussed how data can have a transformational impact on individuals by helping them lead a better life and aiding their decision-making. We have also talked about some of the paradoxes that individuals face in the world of data and how to deal with them. Now, let's broaden our perspective to see how data can benefit our global society. In today's interconnected world, driven by Big Data and digital technologies, we have generated and are generating vast amounts of data globally. This data can be a force for positive change worldwide. With this data, we can collaboratively address global challenges and make the most of the shared opportunities.

By using insights and knowledge from data, we can come together across borders, bridge gaps in understanding and innovate solutions for some of the most complicated issues facing humanity. The transformative power of data goes beyond individuals or nations, crossing geographical borders and nurturing a sense of shared responsibility for the well-being of our world. This journey into a data-driven global society offers us exciting possibilities ahead.

We have seen enough examples where data sharing and collaborations have helped solve some of the most urgent and critical problems of the world. Had the global communities of

scientists and governments not collaborated, the Covid-19 pandemic would have lasted much longer, causing a lot more destruction. Global data sharing and collaboration were instrumental to collectively limit the impact of the pandemic on humanity. Similar collaborations have happened when solving for issues around global terrorism and money-laundering, human-trafficking and more. Whether via bilateral agreements or international organizations, nations have collaborated to solve problems that matter to all.

Collaborations is necessary for transformational global impact

Addressing major global challenges demands a multidimensional approach to understand the complexity of these issues. The largest challenges that we face as humanity are unlikely to be solved by any single nation. Collaborations will result in delivering better outcomes, and global data collaboration is the key to making this happen. It allows us to harness the collective global wisdom and data, to create effective solutions for these challenges.

During Covid-19, the interests of all the stakeholders were aligned and all the stakeholders had a critical role to play in finding a solution, hence we saw effective collaboration globally. Some other examples from world history—where collaboration has formed the basis of driving the future of humanity—have been the Human Genome Project (HGP) that sequenced the DNA of the human race in an effort to know our species better. Another key example is the spread of the world wide web (www).

The world wide web—which started as an experiment for CERN,^{*} and became the bedrock of the digitization of the globe—was an outcome of collaborative efforts of a few global thinkers. Developed by CERN in the early 1990s, it was decided by the leaders of the initiative to make it accessible to all with a global governing consortium—W3C (World Wide Web Consortium). Today, there are more than 400 members globally of W3C, which helps in shaping the global standard for internet, strategies on ensuring internet is

accessible to all, without discrimination, as it is key to building a better world. It was the agreement to make www accessible to all, and not limit it to a few institutions that has transformed the world into what we see today.¹ It was a technology which was developed in a few labs and used for university communication that, when it reached people, created a revolution—the digital revolution.

Both of these initiatives required global collaboration, where visionary thinkers came together to drive transformational change. And they weren't merely reactive responses to crises.

As discussed above, there are multiple examples of global data collaboration, which I will discuss in detail in a separate section ahead, to show how they helped solve critical global issues and benefited humanity tremendously. However, I believe this is just the beginning. As data continues to explode, the opportunities for global data collaboration will continue to increase. Global data collaboration can allow researchers, scientists and policymakers from around the world to pool their data resources, expertise and insights to accelerate research and drive innovation to solve critical global problems that no one institution or nation might be able to solve independently.

Several roadblocks lie ahead

Despite the significance of these achievements and the global opportunities in a data-driven world, several challenges obstruct the widespread scaling of data collaboration at a global level. When it comes to global collaboration, there is often scepticism and protectionism. A fear that the other party may use the information or data that I provide, that is proprietary to me, against me. There is an intrinsic absence of trust.

This makes me wonder whether our world will collaborate to address significant challenges like climate change—which currently do not have an obvious economic incentive? There is a rising awareness about issues like climate change; the UN has adopted the Sustainable Development Goals (SDG) charter, to take the world

forward on preserving the planet.² Having said that, it has still not achieved momentum. There are global bodies advocating the issue, but they are largely ineffective. I wonder how we can enhance their effectiveness. These urgent issues do not seem to garner equal urgency from all stakeholders.

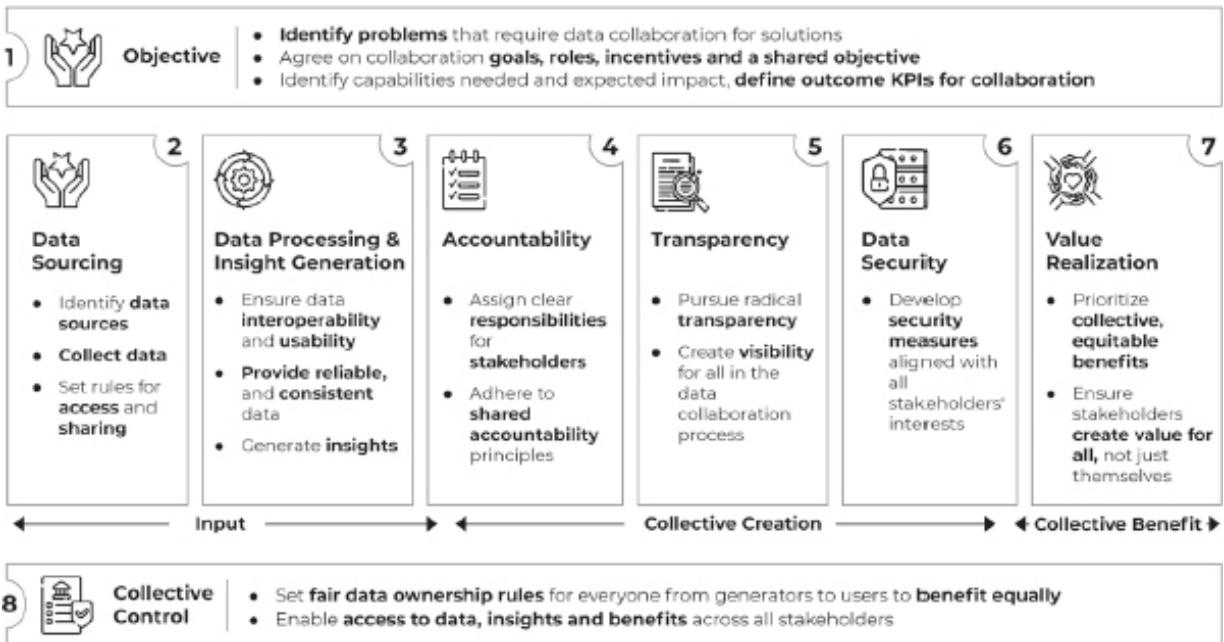
A framework for enabling global data collaboration for a better world

In any global impact initiative, the government and the individual are key stakeholders—the enterprise may or may not participate. The role of an individual is often about contributing to this data ecosystem by sharing their data. They benefit from the policies and services, and co-create the same, by sharing their data. They share their voice, track performance and influence governments and organizations. The governments, on the other hand, are responsible for building the data ecosystem. They monitor progress and evaluate the effectiveness of policies and programmes.

While the benefits of global data collaboration are undeniable, making it happen successfully is easier said than done. There are multiple challenges that need to be overcome, from data privacy and security to data governance and ownership, data quality and standardization, technical infrastructure and interoperability, legal and regulatory complexities to name just a few!

A structured framework is needed to facilitate successful global data collaboration, and I believe it needs to have the following eight step process:

Framework for global data collaboration



It starts by clearly aligning on the **objectives**. The objective of collaboration is to be able to solve large-scale high-impact problems which impact the larger world. Once the objectives are aligned, it is important that the **data sourcing** is thought through, from the sources to the mechanics of sourcing, like formats, channels and structures. Each step is critical to build the foundation for collaboration. The data now needs to be **processed** to generate insights. To do this, the data needs to have interoperability, that is, it must function together seamlessly and needs to be trustworthy and reliable.

Each of the stakeholders must take **accountability** of the data they generate, ensuring its quality and keeping it up to date. The system, on the other hand, has to offer **transparency** to its stakeholders, on how the data is being used, by whom the insights that are being generated, etc. As data is generated via multiple sources, **data security** continues to be a pressing issue. Policies, processes and access guidelines need to be clearly laid out at the get-go.

The **value realization** of the data is when it delivers collective benefits. When it creates a level playing field for each of the

stakeholders. To ensure democratization and prosperity for all, it is important to understand the dimensions of data ownership which includes considerations such as who owns the data, and who benefits from it. Clarity around who owns the data, the governing principles, collective responsibility and **collective control** is key. This is to ensure that people who share the data also benefit and have a voice in how the data is used. This will help remove the 'fear' around data sharing and that we do not end up with a situation of 'data colonization', where powerful entities or organizations from one country or region exploit or dominate the data resources of another country or region.

The one principle that will make any collaboration work well is the theory of 'give-get'. One has to be open to giving—sharing of data—if they want to get the benefits. In simple words, the mindset of 'collective benefits' vs 'individual gain'. To me, collaboration is all about delivering collective benefits via collective responsibility while ensuring collective control. This is important so that no one is left behind—the data generator gets equal benefit from the process as data users and processors.

A governing body: Key enabler for global collaboration

Time and again, we have seen that there is always a global authority that plays a key enabler role towards global collaboration. From CERN and W3C to GenBank (Human Genome Project), to the emergence of GPAI (For AI) (I will discuss these later in the chapter), every collaboration effort required a well-represented global governance organization. I do believe that this need will continue and may accelerate, as data collaboration will only increase. The future, as I see it, will be anchored on collaborations, where data generators will ask for and get a seat at the table and will likely exert control. They will ask for benefits, in return for the contribution they are making. Just like www, data will become the great equalizer and will play its role in democratizing and maybe ushering in a new world order. And the role of these governing bodies will be key.

To be effective, these organizations need to operate based on principles that prioritize **collective benefits**, involve **collective creation**, uphold **collective responsibility** and exercise **collective control**.

While there are governing bodies facilitating data collaboration in specific areas, I believe it will be helpful to have a single governing body that can establish a broad framework and universal guidelines for global data collaborations across areas. The Data Governance Initiative by the World Economic Forum (WEF) is a positive step in this direction, but more needs to be done, given the opportunities and potential growth of initiatives around global data collaboration.

Global collaboration leading to a better world: Case examples

Covid-19, Human Genome Project and massive open online courses and MOOCs are examples of successful global collaborations creating transformational impact. As I lay out the nuances, it is important to understand that these were successful as the alignment around the objective of collective good was understood and agreed by all. None of these initiatives was designed to benefit one over the other, and hence they stand out as discussion-worthy examples.

Covid-19: Global collaboration

In our lifetimes, Covid-19 was possibly one such pandemic which unified the world to fight against a common problem—a virus. Scientists, medical professionals, governments, citizens and more from all over the world, although caught by surprise, openly collaborated, on such a large scale, for the first time, to fight for a common cause. Data was shared and researched during Covid-19 through the use of common data repositories such as OxCOVID19 Database which had epidemiological information on Covid-19.³

One of the best examples of global data collaboration to solve a common problem was identification of the beta version of the virus.

Covid-19 first surfaced in December 2019, and in March 2020, it was declared as a global pandemic by the World Health Organization, WHO. In May 2020, the first variant of Covid-19—beta—was detected in South Africa. This was significant, as this was the first time that the global science community realized that the virus was mutating. This was also the first instance of open global collaboration with South Africa proactively sharing information, data and insights for the collective benefit of the world.

South Africa—which had been sequencing SARS-CoV-2 genomes since the early days of the pandemic—was able to detect the new variant faster than anyone in the world. This was because South Africa was following a sample approach as opposed to a census approach of the UK or other countries. South Africa was being smart about the samples and patterns that it was sequencing and as it saw some variations, as doctors and researchers went deep into that data. South Africa sequenced only 0.8 per cent cases vs 13 per cent of the total cases sequenced in the UK and 4 per cent of the total reported cases in US.⁴ South Africa was able to pull out the anomaly, as they were reading the data not just from their own country, but also those shared by the global community and recognizing patterns and willing to make early guesses. They shared this data with the UK and other nations of the world, who then looked at their data. The strain was identified in South Africa in November 2021, and was classified as a ‘variant of concern’.⁵ This early alert allowed scientists to develop gene-based vaccines rapidly.

The first official cases of Covid-19 were recorded in December 2019 and the first vaccine was approved by US FDA for emergency use in December 2020, which is almost a year from the detection of the virus.⁶ Creating a vaccine in under one year is no small feat. The coronavirus pandemic spurred a global cooperation in vaccine research and development. Even the Ebola vaccine, which was fast-tracked, took five years to reach widespread trials. So what did they do differently?

Scientists use virus genome data to create a blueprint of select antigens. The blueprint is made of DNA or RNA—molecules that hold

genetic instructions. The researchers then inject the DNA or RNA into human cells, which prompts the human cell system to produce antigens.⁷ Thanks to advances in genetics, global collaboration and innovative vaccine development, it became the fastest vaccine development in history, potentially saving millions of lives.

Another noteworthy aspect is how the drug-related data was shared across countries during the Covid-19 pandemic. A culture of publicly sharing draft version of research papers online, which may not have been peer-reviewed, started surfacing due to the urgency of research. This approach relied on the faith that the research community was dedicated to the greater good. A staggering 19,389 articles about Covid-19 were shared in the first four months of the pandemic, a third of which were preprints, unvetted and unfiltered, for all to see, which contributed to the development of the vaccine.⁸

European Clinical Research Infrastructure Network (ECRIN), a consortium of government and public research institutions across multiple European countries, developed a clinical research metadata repository. This repository provided access to the data collected and aggregated in a searchable form to improve accessibility of clinical studies and their related data (example, publications, study protocol, individual participant data, data management plan, statistical analysis plan, web pages, social media).⁹ ECRIN has also developed a toolbox for data sharing, including templates for standard operating procedures (SOPs) and data sharing agreements.

Covid-19 required scientists, governments, individuals to come together, openly share information, be accountable around the quality of what they shared, ensure the data was secure and not reaching any 'not in the know' individuals and yet be in a structure that could deliver swift impact. It is a great example of collaboration for a better world.

Covid-19: Individual innovation

The role of an individual is often about sharing data and collaborating with the system to deliver large-scale impact. During

Covid, it was the individual data shared via smartwatches that helped researchers from University of California, San Francisco, (they tracked changes in heart rate, respiratory rate and other physiological data) to detect onset of Covid-19 and spread of disease.^{[10](#)} These findings were used to develop early warning systems for Covid-19. These models of early detection were shared around the world, and people with smart watches were self-monitoring. This is an example that brings together the idea of individuals collaborating with the ecosystem to solve a larger scale problem.

Education: Example of global collaboration, MOOC-MIT ^{[11](#)}

The first MOOC was designed by Stephen Downes and George Siemens in 2008. Their intention was to exploit the possibility for interactions between a wide variety of participants made possible by online tools so as to provide a richer learning environment than traditional tools would allow. While twenty-five students attended this course in person at the University of Manitoba, an additional 2300 individuals from across the globe participated online.

MIT developed the MITx platform for offering MOOCs. Later, when MIT partnered with Harvard, it was renamed as edX. The non-profit edX consortium which develops and offers MOOCs now has over thirty university partners.^{[12](#)} The consortium has made available an open-source version of the platform, which can be used and developed by other institutions and individuals. Today, MOOCs have emerged as a distinct ecosystem for online learning. More than 100 million students registered on Coursera (world's largest massive open online course provider), enrolled across multiple MOOCs on the platform;^{[13](#)} and multiple players around the world are emerging as unicorns, teaching and collecting data of millions of students.

With education data of over 4 million students, MIT edX has valuable data—from across the world, to understand and improve the quality of education. Concepts like predictive learning, which are

being implemented today, have their roots in MOOC, an evolution that started with an idea of collaboration. Imagine, if all the private players—like Coursera, Khan Academy, Udemy, Udacity, Unacademy, Byju's, Masterclass, Edmodo and more—join forces to share their data and collaborate, what rich information will be available to develop tools that will accelerate and deliver better learning outcomes around the world.

Human Genome Project (HGP)

The Human Genome Project, a publicly funded initiative was launched with the ambitious goal of identifying and cataloguing all 19,000–20,000 human genes. The project aimed to make this genetic information readily available through user-friendly databases like GenBank, fostering further biological research. This project generated a lot of human genome data through the 1990s, but lack of clarity on ownership and collaboration hindered its potential for creating a substantial impact.

However, in a landmark judgment on 13 June 2013, the Supreme Court of the United States in Association for Molecular Pathology v. Myriad Genetics, Inc. unanimously ruled that naturally occurring DNA sequences cannot be patented.¹⁴ The judgment prioritized the collective benefit for humanity over the individual interests of Myriad Genetics. This critical decision rendered genome information non-patentable, opening the floodgates for widespread accessibility to Human Genome Project data. This has implications for many countries and institutes around the world—including India and for many tests including the popular BRCA test, which tests for breast and ovarian cancer. Many countries were able to provide affordable tests as soon as mutations were known, only because the DNA sequences were made open.

Interestingly, as these tests became popular, more data was collected, leading to greater innovation and knowledge generation in the field. This also added commercial value on technological advancement of next generation sequencing (NGS) and sample

processing. This opened up opportunities to create AI-based composite predictors from multi-omics (unified snapshot of the biological processes within a cell or tissue) data.

Another landmark ruling on GenBank and its ownership by the US supreme court stated that just because you have created a database, does not mean you own the data. The data belongs to the people. The Bermuda Agreement, an international agreement among major sequencing groups, mandated that sequence data be shared in a public database within twenty-four hours of generation. This made the data un-patentable but supported scientific progress for the benefit of humanity, promoting fairness.

Both Covid-19 collaboration and HGP collaboration highlight the importance of data sharing. They show that incentives can be aligned to create a level playing field for all, and foster prosperity as long as the focus remains on collective benefits, and ensure there is data solidarity, that is, collective responsibility and collective control.

After discussing the significance and instances of global data collaboration, it's crucial to note that data isn't evenly spread across the globe. Some countries possess abundant data, while others excel in technology and knowledge to leverage this data effectively. The coming together of these data-rich nations and technology innovators holds great potential in unlocking the advantages of data collaboration in the era of data-first world and AI. Let's explore it further.

Data generators collaborating with tech creators: Rise of a new world order

We have seen the benefits of global collaboration that have revolutionized the world we live in. With AI, we will see the emergence of collaborative intelligence and crowd-based collaborations. Just like blockchain, this will be a distributed system of multi-agents, where each agent, human or machine, is uniquely positioned, with autonomy to contribute to a problem-solving network. Design, data and innovation will no longer be siloed and

can be democratized between different stakeholders. However, training the AI will need data, the data that is distributed. There will be locations in the world which will be data generators and there will be locations, which will be technology creators. And it will be critical to see a collaboration between them if significant impact is to be achieved.

For example, today the USA has become the primary hub for AI development—with companies like Google and Microsoft leading the way in creating AI foundational models. China is investing \$27 billion in AI by 2026.^{[15](#)} Japan is working on a society 5.0 initiative, to become AI-ready.^{[16](#)} Korea is investing 20 trillion Won (~\$16.4 billion), towards AI development.^{[17](#)}

That said, India—home to almost a fifth of the world's population—has great potential to become the data generation powerhouse of the world. India's Aadhaar database currently holds 1.35 billion entries with basic personally identifiable information, PII of individuals.^{[18](#)}

India and China, being the most populous countries of the world, hold a unique position in the development of AI and it will be interesting to see the role they play in the development of AI. China has often tended to have a limited or selective approach to global collaborations. India, on the other hand, has the opportunity to be a Big Data generator economy that the technology leaders will need for training their AI models and make them successful. This recognition is evident in India's leadership role in the Global Partnership on Artificial Intelligence (GPAI), a global initiative with thirty-three participating countries.^{[19](#)} GPAI aims to promote responsible and human-centric AI development and usage, underlining the importance of India's role in the AI landscape.

The above example feels like a déjà vu—similar to how the world wide web, ushered in a new world order, AI is creating a seat on the table for countries like India, which is critical to this next phase of global progress. Collaboration is taking the centre stage now to create a better world with data.

Key takeaways

- When it comes to major challenges that we face as humanity, like Covid-19 and climate change, no single nation can solve these challenges alone. Solving them requires global collaboration, and the exponential increase in data offers an opportunity to enable that.
- Despite the number of global opportunities, the widespread scaling of data collaboration faces significant obstacles, primarily due to scepticism, protectionism, fundamental lack of trust and lack of alignment on objectives of the collaboration.
- Effective global collaboration must be rooted in principles such as 'give-get' (or reciprocity), emphasizing collective benefits over individual gains and maintaining 'collective control'. These principles collectively promote equitable value creation for all stakeholders involved.
- A structured approach for driving successful global data collaboration which can safeguard against 'data colonization', ensuring that powerful entities do not exploit another's data resources while enabling trust and easing concerns about sharing data.
- While there are governing bodies facilitating data collaboration in specific areas, it will be helpful to have a single governing body that can establish a broad framework and universal guidelines for global data collaborations across areas.
- The marriage between data-rich nations like India, and technology innovators such as the US, holds immense potential in unlocking the advantages of data collaboration within the data-first world and the era of artificial intelligence.

Data as a Source of National Competitive Advantage

Twenty-First-Century National Asset

'This nation is never finished, it is recreated by each generation'

—Lee Hamilton,
American politician and lawyer

With each new generation, global dynamics evolve and the world order shifts. And each generation creates its own competitive advantage. As we read history, we can see the sources of national competitive advantage have evolved over time. In ancient civilizations, trade centres and knowledge hubs were a source of competitive advantage and had helped civilizations thrive. Fast forward to the era of the Roman Empire, their competitive advantage lay in powerful legions and weaponry like catapults which allowed them to conquer vast territories and maintain control over their empire. Global power dynamics shifted again as heavy artillery and cannons emerged as a source of competitive advantage. In the age of exploration and multi-continental empires, the British Empire seized the limelight. Their competitive advantage was derived from the extensive use of gunpowder and their naval supremacy. With the UN, when the world reached some stability (although not complete), power displays continued with deadlier weapons—from nuclear to bio-warfare. We are now in the digital age, and this generation of digital natives and digital disruptors is recreating sources of national competitive advantage. Nations are also redefining their roles and trying to identify the new sources of national competitive advantage.

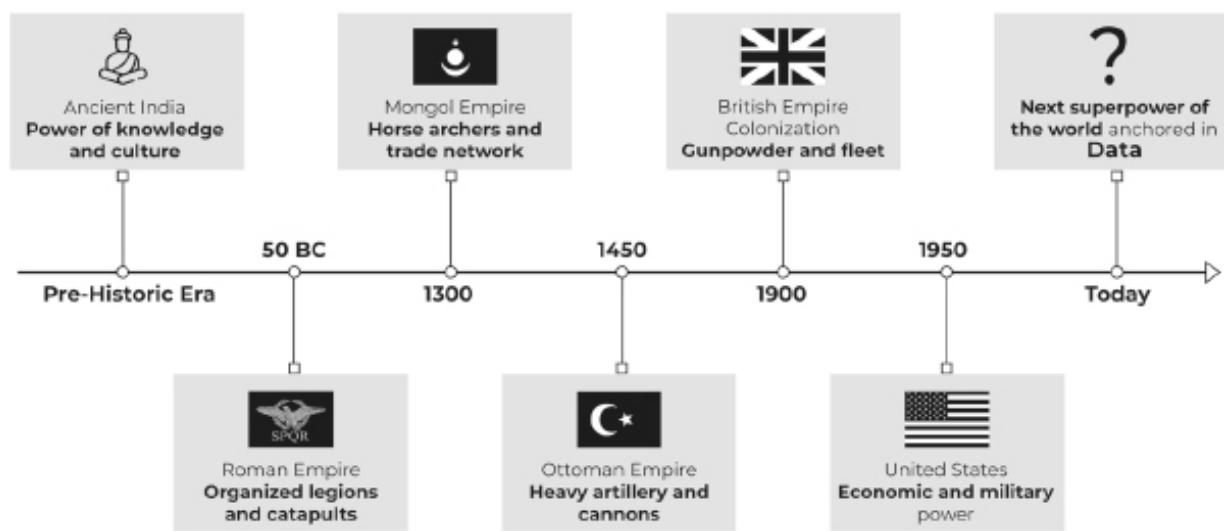
In my lifetime, I have witnessed the era where the US and USSR were leading as the world's superpowers, competing intensely, building up their military, creating atomic weapons and racing to conquer outer space to prove their dominance. I have seen the falling apart of the USSR, the meteoric rise of China and to an extent India, and now, in the age of the digital and data-driven world, the rise of Israel, South Korea and Estonia as emerging centres of excellence. These countries may not be considered as global superpowers, but they are certainly gaining a competitive advantage in our data-driven world that many other nations are still grappling with.

In the last chapter, I highlighted the importance of global data collaboration for addressing some of the most difficult problems we are facing as a society. But at the same time, data can also serve as a source of powerful competitive differentiation for nations at a global level. This is similar to the concept of multi-source and proprietary data that I have talked about in Section II. While data from multiple nations can help generate deeper insights and solve problems which no nation can solve alone. Data can also serve as proprietary knowledge to build differentiated capabilities, setting them apart and giving them a competitive edge over other nations.

As discussed in the last section, the landscape of competitive advantage for organizations has undergone significant transformation over the years. While the previous industrial age primarily emphasized machinery, land and factories as core differentiators, the focus gradually shifted to leveraging technology for gaining a competitive edge. As a result, every company in the digital age started focusing on becoming a technology company. However, in the data-first world, enterprises are focusing on becoming data-first organizations and data is becoming a key source of differentiation. It is evident that the profound impact of data extends not only to the realm of enterprise operations, but also holds immense potential for nations to distinguish themselves through data-driven strategies. We have seen a number of examples of how data enables brands and enterprises to shape consumer behaviour and preferences, and this potential extends to nations as

well. By harnessing data, nations can cultivate soft power by influencing the behaviours and preferences of individuals and assert their influence on the global stage. More importantly, it can also enable them to deliver enhanced services and prosperity to their citizens. It is now clear that data will emerge as a potent source of competitive advantage for nations, ultimately determining the next global superpower.

Source of national competitive advantage has evolved over the years



Data as an enabler for prosperity, hard power and soft power

Data has the capability to bring prosperity, hard power and, soft power to a nation, which can create a competitive differentiation. Let me explain this:

- 1. Prosperity:** Prosperity is a measure of a nation's wealth and quality of life. Data can help enable better services such as e-services that enhance government-citizen interactions and boost efficiency. Nations can leverage data to advance Smart Cities, real-time monitoring and maintenance of infrastructure, smart mobility, energy management and safety measures. Data-driven decision-

making also enhances healthcare and education systems for citizens. Data helps national decision-makers to craft policies which can deliver desired economic outcomes, enable wealth creation and improve the quality of life for citizens. For example, Singapore leveraged the potential of data to evolve from a trading port into one of the world's most reputable commercial hubs over the span of six decades. Even seemingly simple measures, such as enhancing English language proficiency of students through data-driven monitoring by the ministry of education,¹ have contributed to making Singaporeans a highly skilled and globally competitive workforce. The data insights driving decisions on what skills to prioritize in education, the adoption of adaptive learning methods and tracking progress are all firmly rooted in data. This exemplifies the integral role data plays in Singapore's success story.

2. **Hard (military) power:** Military power historically was about weapons like nuclear weapons, air and naval strike capability, etc. Data plays a vital role in military hardware development, aiding in design optimization, performance evaluation under various conditions and the improvement of training programmes for military personnel. Countries like China are investing significantly in protecting as well as creating technologies to establish leadership in autonomous weapons.² The war between Ukraine and Russia is being fought by the autonomous Turkish Bayraktar drones.³ The popularity of the drones is exceptional that the Ukrainian army has a song —‘Bayraktar’—celebrating the drones. I would imagine, in the very near future, cyber warfare is going to be another key determinant of a nation’s power of defence. Technologies such as ML and AI enable the analysis of extensive data for pattern detection, anomaly identification and predictive modelling. In cyber warfare,

these techniques enhance threat identification, attack response and defensive measures' effectiveness. Utilizing data, security systems can enhance their capabilities by collecting and analysing network and system data to swiftly respond to potential threats.

3. **Soft Power:** Anyone who controls the media, controls the public narrative. From Donald Trump to Jeff Bezos to Elon Musk, everyone has media interests. And anyone who can control the public narrative can influence public opinions and sentiments. It reminds me of the quote by Jim Morrison: '*Whoever controls the media, controls the mind.*' From print media, now we have social media. Data has the power to shape global cultures and behaviours. By examining data related to cultural preferences and behaviours in other countries, governments can pinpoint common interests and cultivate a greater appreciation and influence of their own culture on the international stage. Imagine the power when a country starts controlling the narrative in another country by way of cultural influence. K-Pop and K-Drama are taking over the world. BTS, a popular K-Pop band, with its 'ARMY' across the world is possibly one of the biggest influential mobilizers in the world, and they are dependent on data.⁴ They use data from its X (formerly Twitter) and YouTube followers to define the content development and drive the mindset. They use data to understand the preferences and thoughts of its followers around the world and leverage it to make new content, and then influence the actions of people around the world. This, to me, is an unmistakable sign of soft power of data in controlling the world.

The data race is real, and it is just the beginning. There are no clear winners currently, but there are some early leaders like China, South Korea, Estonia, Israel, which may have left behind nations like Russia in this race for data dominance. South Korea with its smart cities

(example detailed out later in the chapter), Estonia with its national ID system which integrates taxation, banking, health and voting—the key pillars of any citizen services in a nation, have proven that it is possible to deliver prosperity, hard power and soft power with data as a key enabler.

These countries have been strategic about how they have developed and leveraged data ecosystems and taken advantage of these to build better citizen services, and this is just the beginning. They have laid the foundation to potentially gain a competitive edge and set global benchmarks for quality of life and success metrics. Let us explore some of these examples.

Harnessing the power of data for economic and competitive advantage: Case examples

When a villager in India is able to verify their land records on the internet on a government portal, we know the country is making progress on the path to development. The potential of a 24x7 accessible database which has details of each and every landowner of a nation, is a testament to the power of information and data. Imagine the tremendous power and possibilities of using this data to deliver benefits.

As technology becomes more sophisticated, the ability of nations to generate and analyse data improves massively. Countries need to invest in building data harnessing capabilities (supported by policy) to deliver benefits for its citizens. Countries who have been successfully able to harness the data, have transformed it into a competitive advantage and a valuable trade asset, establishing them as sector leaders at the global level.

Estonia became the first model digital nation of the world in 1994

In 1994, a small community of government officials, IT specialists and academics developed the strategy paper 'The Estonian Way to

the Information Society' with the aim of establishing principles for the management of modern, efficient state information systems.⁵

The goal was to solve social challenges stemming from political uncertainty with IT solutions. To be successful, they developed an engagement model where they collaborated with the industry, taking citizen feedback and improving on solutions. They created working groups focused on the effort with access to the cabinet. A model that they feel so confident of working that they are leading the world and are open to share with anyone interested.

By 2001, as the nation was getting more internet savvy and data was increasing, Estonia launched an x-roads initiative, which integrated data platforms, to reduce data exchange costs and prevent data leaks from unsecured platforms. As a result, 99 per cent of the public services were made accessible online 24x7.⁶ E-id and e-signatures across all services were introduced in 2002,⁷ which saved the country 2 per cent in GDP,⁸ besides driving efficiency. In 2005 they introduced I-VOTING. In 2008, Estonian cryptographers developed and implemented block chain for government registries (a technology, which is still experimental in 2023 in many countries), in an attempt to create another secure layer to prevent harms from cyberattacks.⁹ By 2010, 99 per cent of citizens had e-prescriptions.¹⁰ In 2015, Estonia created the first data embassy—a high security data centre outside of Estonia as disaster management.¹¹

In 2019, Estonia started working on policy of government services and AI.¹² Its goal was to create the legal and strategic framework for accelerating AI development, making Estonia a trailblazer in the field. The stated outcome—a detailed strategic plan for promoting implementation of AI solutions in the public and private sectors. Estonia today, with all the data it has about its citizens, is proactive about how it supports its citizens like offering pro-active childcare benefits. Parents of a newborn no longer need to apply for benefits, as the state already has the information. Estonian's claim that their

level of state infrastructure is so sophisticated, that everything can happen online but getting married and divorced where people have to meet. Imagine, buying and selling real-estate online. It is rare that any public service can offer that level of digitization.

I bring this up, because until a decade back, no one knew about Estonia, and today the world is reaching out to the experts in the country to learn from them. Clearly, they have been able to make a mark for themselves on the global map. They have developed a competitive advantage and are taking a potential leading position in the emerging data led world. How long they maintain it will be a function of how much they are able to work across policy, execution, legal, stakeholders and orchestrate the vision to implementation journey. But they surely got it right the first time.

Israel harnessed the potential of AI to pioneer world-class cybersecurity platforms globally, led by their government vision

Tel Aviv harnessed the power of data and AI to develop itself as the cybersecurity capital of the world. Their exceptional capabilities have led even their rivals, Saudi Arabia and the UAE, to become their primary cybersecurity trade partners. \$8.8 billion were invested in 2021 in cybersecurity start-ups in Israel.¹³ There are two companies, Check Point Software Technologies and CyberArk, which have a combined market capitalization of ~\$22 billion, as of 11 October 2023.¹⁴ Tel Aviv is rising, and is second to Silicon Valley, when it comes to cybersecurity start-ups.

Israel was able to harness this power, due to government policies, support and vision. Think about it, in Israel, cybersecurity education starts as early as middle school and Israel is the only country in the world in which cybersecurity is an elective in high school matriculation exams. Israel was the first country in which you could get a PhD in cybersecurity (as an independent discipline, not as a computer science subject). Such achievements were only possible

because of the government support for these initiatives and their vision to build Israel into the cybersecurity capital of the world.

The government also enabled partnerships with the military—it acted as an incubator for cybersecurity. Government and military data sets, combined with technology from private players, accelerated Israel's development as a cybersecurity leader.

With years of intelligence-gathering and cybersecurity practice, the Israel Defense Forces' (IDF's) Unit 8200 has evolved into an incubator and accelerator of Israel's start-ups, in cybersecurity and other fields.¹⁵ 'In the past, military service was perceived as a waste of time, but it is different now. We didn't plan it that way. No one thought about how to make the IDF into a catalyst for the Israeli economy, but that's what happened,' says Nadav Zafrir, former commander of Unit 8200.^{16 *}

The government of Israel played a pivotal role in harnessing the power of data. They dedicated resources, developed talent, leveraged cutting-edge technology, including that used by their military. They fostered innovation, and most importantly, built a thriving market through improved international relations that further facilitated this progress.

South Korea has harnessed the power of AI to improve citizen services

South Korea ranked first among twenty-nine OECD countries in the 2019 Digital Government Index which evaluates digital maturity of government services.¹⁷ This has been achieved through a focused multi-year effort. For instance, Seoul executed its Smart City development in phases, starting with fully online e-governance services (1999–2007), open digital platforms for public information and participation (2007–11) to customized spatial services based on citizens' needs and decisions based on data (2011–present).¹⁸

The city has followed a four-step process: collect, analyse, understand, and problem-solve to launch many services. One of

such services is the night-bus service. Data about taxi ridership and floating population was collected and analysed. Based on the insights, routes and timings were decided, covering 42 per cent of Seoul residents. The service led to a 10 per cent increase in ridership, boasting ~10K rides per day.^{[19](#)} Additionally, the service resulted in an 11 per cent increase in women's activities at night.

Additionally, Seoul city also mitigated road accidents through analysis of transport infrastructure vulnerabilities. Data used included traffic-accident history, floating-population data, weather data and car-speed data. The analysis resulted in identification of black spots, installation of speed bumps and pedestrian infrastructure, leading to a sharp reduction in recorded accidents.

They primarily used public data, including data from central government agencies, IT systems, data from city infrastructure such as CCTV, public transport databases. Data from private entities was also sourced, including mobile-phone records, credit-card data, floating-population/spatial data.

South Korea adopted a very forward-looking approach to use data and technology to improve public services. Their 'Digital Twin Seoul S-Map' project replicates Seoul in 3D within a virtual space, setting a unique global precedent for addressing urban challenges through data. While 3D maps existed before, this marked the first instance in South Korea of constructing a digital twin capable of analysing and simulating city-wide issues such as planning, real-time fire monitoring and wind-pattern optimization.^{[20](#)}

India has harnessed the power of data for Aadhaar, for sharing direct benefits to citizens

India has leapfrogged when it comes to using data as a source of nation-building. The Indian government launched the Aadhaar project in 2009. This project laid the foundation of India's digital transformation and stands as one of the most unique and pioneering endeavours for its ~1.4 billion people. Aadhaar has been a game changer for India. It is an identification number linked to an

individual's biometric data, personal information and address. By December 2022, the cumulative number of Aadhaar authentication transactions had crossed ~88 billion and an average of 70 million transactions per day.²¹ A majority of them are fingerprint-based authentications.

One of the impacts of Aadhaar today is know your customer (KYC) link to JAM. JAM is the acronym for Jan Dhan (a financial inclusion programme of the Government of India, to provide banking services to all Indian citizens), Aadhaar and mobile. To curb the leakage of government subsidies, the government has simply linked the Jan Dhan bank accounts, Aadhaar cards and mobile numbers of Indian citizens and called it the JAM Trinity. JAM is essentially the combination of the above three modes used to deliver direct cash benefits to the bank accounts of the eligible beneficiaries by the governments. This has effectively addressed a major problem for India as direct benefit transfer schemes in India were fairly complicated, or in most cases, intended benefit was not fully reaching the citizens.

Aadhaar facilitates direct beneficiary identification through biometric data whereas Jan Dhan accounts and mobile phones enable the government to transfer cash directly into the bank accounts of the beneficiaries.

Building ability to leverage data as a source of competitive differentiation

I am a firm believer that any successful transformation doesn't solely originate from the top or bottom of an organization. It's a process that involves setting agendas at the top and executing them throughout all levels of an organization. Effective transformation requires collaboration and co-creation among top, middle and bottom levels. This highlights that transformation is not a matter of chance, it must be carefully planned. It involves developing capabilities over time and translating them into tangible impact. In

the realm of data-led transformation and using that as a national competitive advantage, I suggest a four-step process.

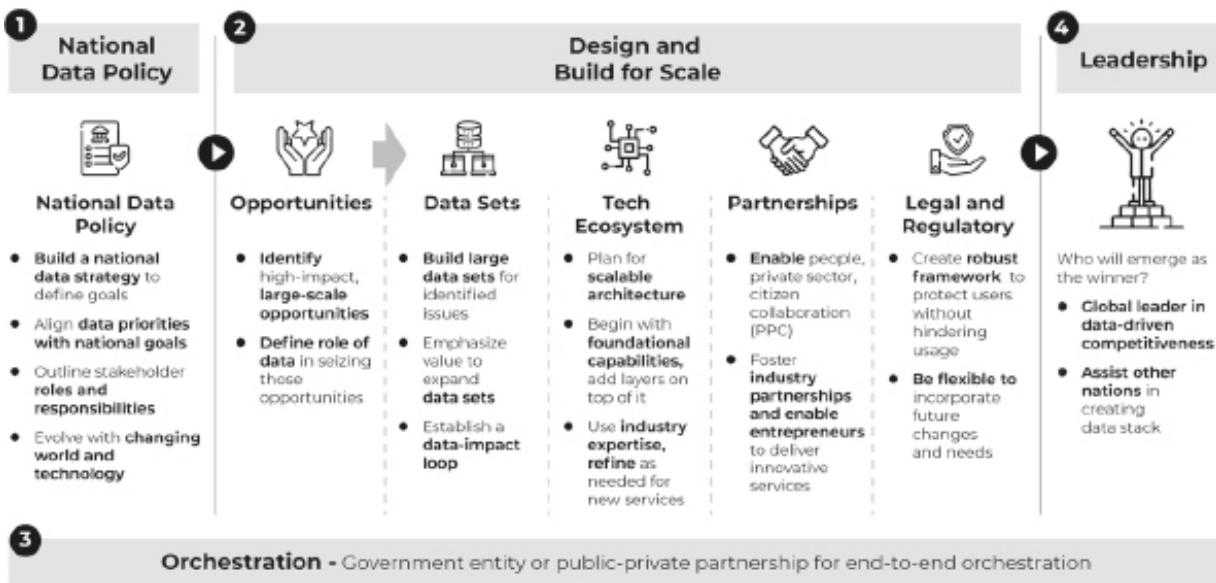
1. National data policy

A national data policy is the starting point for any transformation at scale. Leaders need to envision how their country intends to use data to address the nation's most pressing challenges. The data priorities should be aligned with the national priorities, while giving enough flexibility to make improvements, as technology evolves.

A data policy that is focused, where the government has identified specific opportunities for impact, including use cases and focus sectors—example, healthcare, education, citizen services or more—is more likely to succeed. Estonia focused on delivering prosperity by going digital across health, education and citizen services. In India, Aadhaar aimed to foster prosperity and inclusive growth, with citizen services being the first large-scale visible impact.

A national data policy is a vision document, anchored in principles that can evolve over time and have the strength to deliver long-term value to its citizens. It needs a deep understanding of the national agenda and the expertise of industry and academia, to become relevant. In India, Aadhaar is a classic example where the co-founder of Infosys, Mr Nandan Nilekani, was brought in to lead the Unique Identification Authority of India (UIDAI) effort, along with two IAS (Indian Administrative Services) officers, by making UIDAI into a statutory body under the ministry of information and technology. A similar focus may be needed when countries think of building a national data policy and delivery impact.

Framework for nations to build data as a source of competitive advantage



2. Design and build for scale

Identifying and prioritizing opportunities for large-scale impact, where data is a key enabler is a complex task and is the next step following a policy decision. This needs to be done by relying on the wisdom of experts, who have a deep understanding and experience of working on national priorities. Typically, these opportunities fall in the domains of poverty alleviation, healthcare for all, education for all or better citizen services to deliver a superior quality of life. At the end of it, the objective is to deliver prosperity at large scale.

The execution needs to be designed and built to scale. Choices have to be made on data sets to be collected, the technology backbone to be established, the partnerships that will be the enabler and the governance, with legal and regulatory framework. Without people's participation, any policy will fail in implementation.

Engaging with citizens groups at an early stage is beneficial. In addition to citizens, the government may need to borrow expertise from industry and academia to build large data sets and enabling technologies. Such collaborations and partnerships help build a robust model for both design and execution.

I also believe that when a nation is targeting large-scale transformation, the foundational capabilities are critical. In the case

of data-led transformation, the foundations have to be laid out such that it is scalable at the get-go, and new layers can be added as the system matures. And these need to be resourced—with money and talent at times—sometimes from government budgets, but ideally structured for self-sustainability in the long run.

We also understand that no nation can build the entire ecosystem perfectly in a short time. It will always be staged and implemented in phases. In India, the Aadhaar was first piloted in January 2009. It had the eKYC and was linked to JAM. It was then linked to payments in National Payments Corporation of India (NPCI). Now, in 2023, it is expected to have a layer of Open Network Digital Commerce (ONDC), exhibiting the stacking and the layered approach to building data capabilities.

Being open to changes and having the patience to scale up in stages is important. Active people participation helps. Implementing infrastructure changes should be coupled with awareness campaigns to encourage greater public involvement and accelerate the scaling process. But all of this needs a robust legal framework, an essential element that will be key to protecting what is being generated—an element that often comes as an afterthought but should be envisioned at the outset.

3. Orchestration

While one may devise the data policy and plan for a design and build to scale, when speaking in the context of a nation, the efforts need to be orchestrated across multiple stakeholders and interests.

Achieving alignment across all and being able to drive the agenda, needs strong public-private and citizen partnerships. These partnerships need to be agenda-led and will need an overseeing body that ensures delivery. I have spoken about UIDAI in the case of Aadhaar in India. These organizations should be open to feedback, pre-empt the challenges, proactively plan for the future and remain open to continuous improvement. Achieving all of this requires dedicated focus, and I strongly advocate setting up a public-private

partnership body that will drive data-led prosperity initiatives for any nation.

4. Leadership

Finally, a leadership position is achieved when one starts harnessing the data in a seamless, automated and predictive manner (as exemplified by Estonia) and realizes the target impact. True leadership for nations is about unlocking the transformative power of data in creating a positive impact to improve the lives of citizens and enhance the overall well-being of a nation. Leadership also entails utilizing the positive changes and advancements made possible by data to elevate a nation's position and reputation on the global stage.

The digitization of payments services in India is a great example in this regard. The NPCI is taking its popular digital payment system, the Unified Payments Interface (UPI), to other countries. UPI accounted for 52 per cent of India's total digital transactions in FY22.²² In September 2023 alone, NPCI handled nearly 10 billion UPI transactions, totalling over \$180 billion.²³ India is taking a leadership position in this space and is now sharing the success of this initiative with other nations, holding discussions with thirty countries regarding UPI adoption. France, the UAE and Singapore have already adopted this open digital platform, while Nepal, Japan and China are at various stages of negotiation. NPCI plans to use UPI for cross-border money transfers. In June 2022, NPCI International signed an agreement with France's Lyra Network to accept UPI and RuPay Cards in the country.²⁴

Many nations are able to achieve leadership positions early, the challenge is often holding on to this position. Can any other nation catch up? Why is this leadership position important? Securing a leadership position carries significant advantages, including the ability to exert influence on other nations. It also elevates the nation to a prominent global platform, strengthening strategic and business partnerships, and paving the way for various industrial and economic

prospects. I believe that nations have the opportunity to achieve this if they start recognizing the value of their data and building a vision and capability to harness it.

Who will win? It's too early to say, but the spot is open right now, with no clear winners. USA is currently the front-runner, having the largest and most influential technology companies in the world, along with a formidable educational infrastructure and a thriving innovation ecosystem. But can it hold on to its leadership position, or will China overtake it, especially given the massive investments the latter has been making in AI? Or will China stumble, given its challenges around data democratization, a critical feature and enabler of the digital age? Can India be a dark horse in this new data-first world? I believe that India has the potential and the opportunity to become a leader in the data-first world. Its robust IT industry has already positioned it as a global technology powerhouse. Aadhaar, a groundbreaking initiative, is widely regarded as one of the most capable digital stacks. This has already helped India make significant strides in becoming a digital economy and promoting financial inclusion. As I mentioned in the last chapter, India can also potentially become the data generator for the world, partnering with technology-rich countries.

However, it is not just about India. Could we see some countries from Africa and Latin America breakthrough? Technology and data disruptions are levelling the playing field and creating opportunities for many countries to leapfrog. It will be fascinating to observe how the global power structure shapes up over the next two decades.

Key takeaways

- The source of competitive advantage for nations has evolved over the years. We are already witnessing data emerging as the source of competitive differentiation for nations in the twenty-first century.
- When effectively harnessed, data holds tremendous potential for empowering nations to achieve national

prosperity and exert both hard and soft power on the global stage.

- We have already seen several nations such as Israel, India, South Korea and Estonia that have set exemplary examples of harnessing data to build their competitive differentiation.
- Designing a national data policy, identification of high-impact areas, building large data sets and leveraging ecosystem partnerships are key ingredients for any nation to effectively build capabilities for harnessing data.
- Orchestration anchored in strong public-private-citizen partnerships, driven by a clear goal and overseen by a dedicated governing body is critical for ensuring the success of data-led initiatives.

In Conclusion

Mastering the Data Paradoxes to Win the AI Age

Life is data, data is life

At the beginning of this book, I had laid out three questions I wished to address:

1. *Why and how can AI powered by data create transformational value for enterprises and individuals?*
2. *The world is full of paradoxes and so is the world of data. How can individuals and enterprises effectively deal with the paradoxes, to unlock the transformational value of data and AI?*
3. *Are there some principles that we can learn from life and apply to data and vice versa, as we navigate the new world of data-enriched lives?*

Through the chapters, I have attempted to provide answers to these questions. I now conclude this book by bringing it all together.

By now, as my readers, you would be very aware of my passion for 'data' as a topic, and how I believe data serves as the backbone of the digital age and plays a fundamental role in shaping our lives. The digital age, which has accelerated the growth of data by leaps and bounds, has intertwined data and life, significantly blurring the boundaries of the digital and the real worlds. I believe that today, we live in a world of hybrid reality, which is a mix of digital and real. We are continuously generating data and enjoying the benefits of this data and generating more data in the process. In a way, our life has become data and data has become life. Life has always been driven by data, and in the data-first world we are in, it is even more so. At

the same time, the nature of data mirrors that of life. And as life, data also presents multiple paradoxes, a seemingly impossible-to-solve conundrum. However, if you look deeply, the problems and principles to solve them are not different from what life has to offer.

In my previous book, I focused extensively on the digital age. One of the key things that I talked about there was the advent of AI. AI cannot happen without data. This book highlights the recursive relationship between digital, data and AI—the trinity, that is expanding, growing and transforming the world we live in. I am very excited about these topics, as these are real today and are sparking the future. AI is not just a technology-enhancement, it is a revolution riding on the back of data, and it is happening today; and we all have a role to play in accelerating this.

1. Why and how can AI powered by data create transformational value for enterprises and individuals?

AI has been brewing for many years, in the research labs of many tech giants and the universities. But for any revolution to happen, the timing has to be right—where many things have to come together. But according to me, the most critical development that has happened is the ability to tap into the **wisdom of crowds**.

With the digital age and data abundance, the unlock is happening right now. We are the fortunate generation witnessing this emergence of the AI age. We have an opportunity to participate in it, shape it, benefit from it. AI is no longer happening in isolation. It is happening on every computer, driven by every click we make. With Gen AI, which is built on massive internet data, we now have access to a vast repository of knowledge from a diverse group of individuals—their collective opinions, insights or expertise can be leveraged to arrive at more accurate and insightful decisions or solutions than any single individual could come up with. The giant layer of smart data, which forms the foundation for the neural networks, is akin to the metaphorical concept of **universal**

consciousness—an underlying essence of all being and becoming in the universe. These neural networks are becoming so sophisticated, extensive and interconnected that they are becoming the underlying foundation for knowledge and problem-solving on a vast scale. Gen AI continues to generate more data, creating an infinite loop of learning and refinement, and reinforcing the virtuous cycle between digital, data and AI.

The possibilities are mind-boggling. Never before in the history of humanity has technology connected the world as ONE. Yes, with the internet we had access to the data, but it was left to the individual to search, process and generate insights from it. And with traditional AI, the tools connected the dots, but had limited impact due to limitations around data availability. But with Gen AI, suddenly the constraints have been removed, and now the AI algorithms not only have access to enough and more data to learn, refine and enhance their accuracy, but also to generate new and novel content, enriching the data pool further. As the data will continue to expand, the virtuous cycle of the AI age has just begun. The AI revolution is just beginning and is going to be significantly different from the industrial revolution. Let's see how.

AI is unlocking value across multiple, deeply interconnected levels

While AI is an unprecedented opportunity, the true unlock will happen when the data at the macro-level, which is the massive data set currently being sourced from the internet, starts working in tandem with the narrow and deep enterprise proprietary data and the individual personal data. That's when we will see highly personalized, impactful solutions for individuals, and AI operating efficiently at scale for enterprises, addressing specific problems effectively.

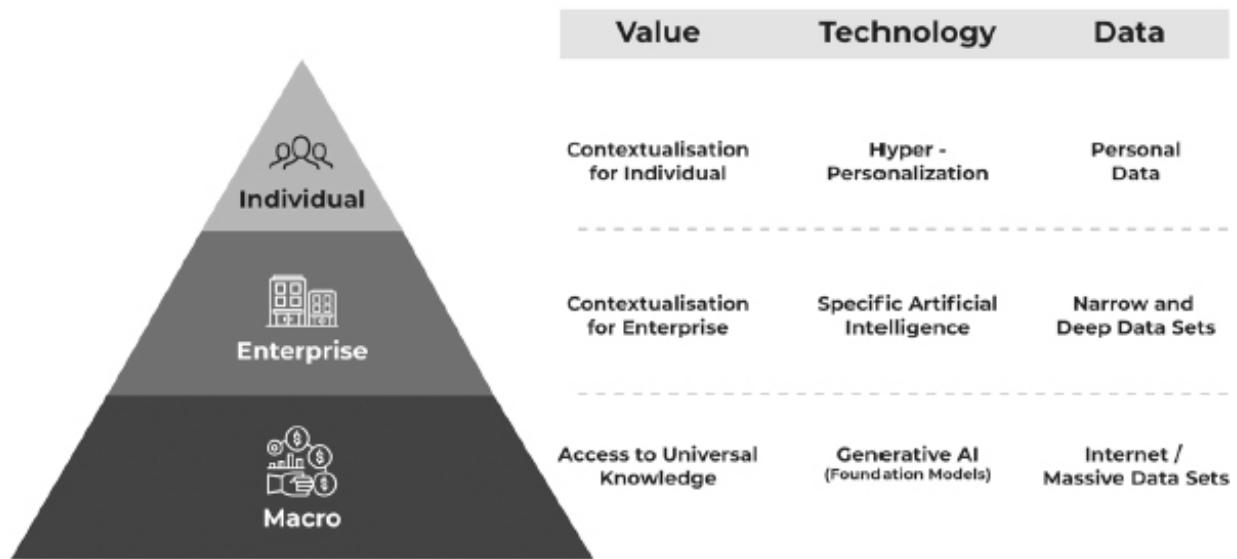
The Macro

Gen AI models like ChatGPT and Bard, etc., are based on LLMs that are exceptionally advanced AI systems that have been trained on extensive data sets, the massive data of the internet. These models provide the necessary foundational framework for solving complex problems by leveraging the wisdom of the crowds. It is our collective data that forms the foundation for the neural networks that are currently evolving and integrating, to form this giant foundational layer that enables us to tap into such an enormous knowledge base, generated and shared globally.

The revolutionary shift here is that LLMs can process information and provide sufficiently accurate and useful solutions to a wide range of difficult problems that previously required significant time, effort and expertise. While they may not complete solutions or be as customized as required to solve specific problems, it can get us as close to 80 per cent of the solution.

Let me give an example. Imagine you are trying to build a new software product. You would typically reach out to experts for their opinions, look at other products, compare and then come up with a feature set. But there are limitations—as you will have a limited outreach and processing power, grasping ability, and of course, time. Gen AI tools have an ability to do all the above steps for you within a few seconds. Now all you need to do is further refine and validate the AI-generated feature set to ensure that the software aligns perfectly with the needs and preferences of your target audience.

Generative AI - Unlocking the transformational value by connecting macro, enterprise and individual layers



The best minds, the best products, knowledge is today behind these tools, and the technology is intelligent in processing and matching their response to your questions. Now, the onus is on you, how smart a question you can ask? The paradigm of man vs machine is now shifted to man and machine, where the machine will accelerate your speed of thinking, as it has developed an exceptional speed of high-quality processing, anchored in data of the world.

The enterprise

AI in its foundational form is able to make sense of all the data in the public domain. Like I said, that's 80 per cent of the job done. Now the critical 20 per cent, which creates a unique differentiation of the organization, can only be achieved when the foundational LLM model is layered with enterprise-specific proprietary data. This would enable the organization to solve specific problems more comprehensively, enabling it to move the solutions from conceptual ideas to practical applications to accurately meet the organization's unique business requirements.

This requires specific AI models trained on enterprise data sets. While the current focus is on enhancing AI for efficiency and

proficiency, I believe AI will soon transition beyond proficiency and efficiency into solution-driven innovation at scale, enabling organizations to explore creative solutions quickly. To achieve this transformation, they must adapt their culture and talent. Data, and the value it generates, must become a primary focus for enterprises to maximize AI benefits, as AI underpins the delivery of high-quality, highly personalized customer experiences, fundamental changes in business operations through automation and AI workflows, and the emergence of innovative AI-driven business models.

Gen AI, coupled with the meta-layer of foundational models fuelled by internet data, is democratizing the data-driven landscape. Traditional industries like manufacturing and healthcare, which were not inherently data-first, can now harness this metadata layer to drive insights and potentially transform their business models. This accessibility to a wealth of data will encourage greater participation and innovation, breaking down data limitations for less technologically advanced enterprises.

The individual

In Section III, Chapters 21 to 25, I focus on the individual implications of data and AI. At an individual level, the trend of hyper-personalization will continue, with large volumes of deeply personal data being generated. Right now, individuals are using Gen AI for curiosity and experimentation. The impact one is seeing is minuscule, but the processes behind it are smart. For example, if someone used a Gen AI tool to auto-generate copy for an advertisement campaign, the tool will be sourcing data from multiple other copies, mapping the trends and then delivering a copy based on the prompt that individual has given. The next phase will be when AI is further contextualized at an individual level, to solve specific individual problems to deliver specific impact. From healthcare, to finances, to relationships, AI may be able to find personalized solutions to all individual problems, by understanding the specific individual needs, as learnt from the individual data

shared, and the most successful solution available in the given circumstances, based on the data from wisdom of crowds.

The three layers are reflective of a deeper reality!

The interplay of the three layers—the macro, the enterprise and the individual—that I have talked about above is deeply powerful and has the potential to create unprecedented value. These data and AI layers that we see emerging have very interesting parallels from the construct of the mind/self as articulated in the famous Indian text, the Yoga Sutras of Patanjali. Yoga Sutras lay out four layers—*Manas* (the thinking mind), *Ahamkara* (the ego or the identity), *Buddhi* (the intellect) and *Chitta* (the universal consciousness). Most times, we are operating from *Manas* or the ‘small mind’. Dramatically superior intuition and intelligence can be accessed if we can connect to the deeper layers of self. Something very similar is happening with Gen AI. Individuals and enterprises now have the ability with Gen AI to access the ‘wisdom of crowds’, the ‘universal consciousness’, which dramatically increases possibilities for impact, and I believe this a revolutionary unlock for human civilization.

AI owes its acceleration to data abundance

What has suddenly changed in the recent past, to trigger such an explosion in the realm of AI? The answer is data.

For AI to be effective, the triad of data, computational power and algorithms is needed. While the latter two were being taken care of to a large extent, the differentiator, and the hardest problem to solve for a long time was to find large volumes of data to train the models. Lo and behold! We are there now as data abundance has enabled Gen AI. AI has moved from a potent and effective technology to a transformational one.

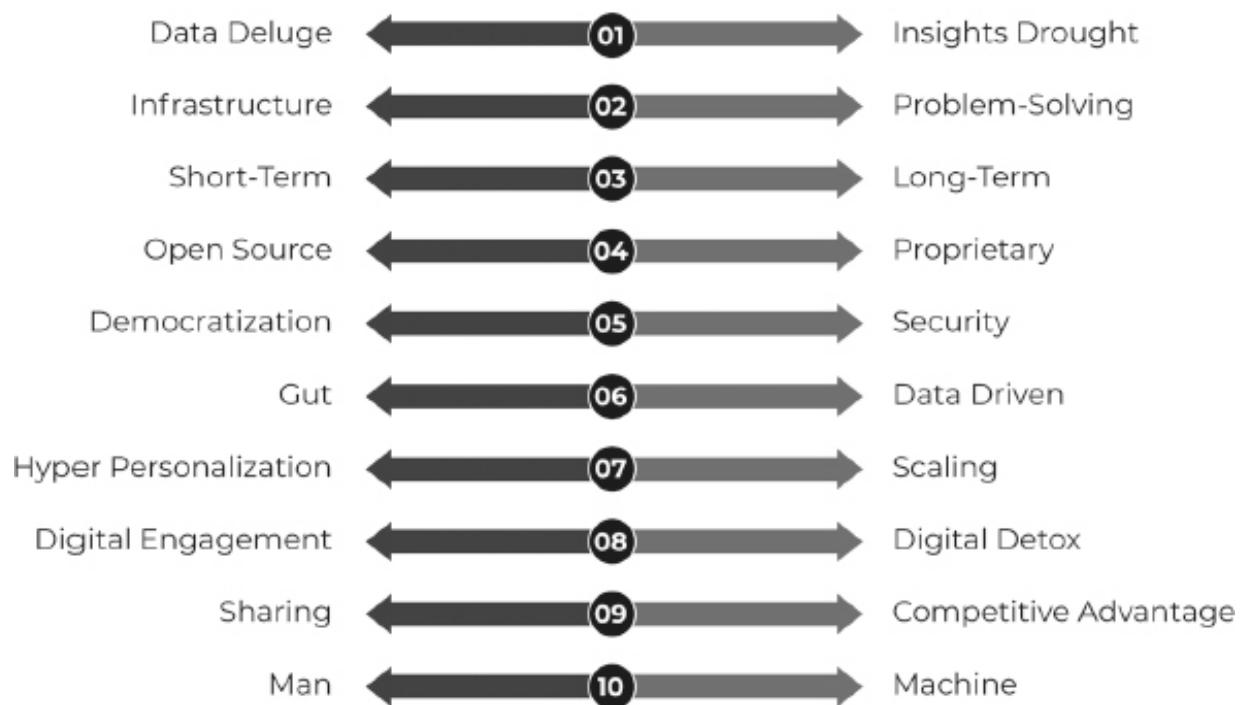
But the data world has its own set of challenges and paradoxes. As I have mentioned in the introduction, the way the AI age unfolds hinges on how well we are able to address the data challenges and paradoxes. We need to understand these paradoxes and find ways

to tackle them. I lay out the data paradoxes and my ten principles to solve these.

2. The world is full of paradoxes and so is the world of data. How can individuals and enterprises effectively deal with the paradoxes to unlock the transformational value of data and AI?

Like life, the world of data is also full of paradoxes. Across Sections I, II and III, I have laid out multiple paradoxes that emerge as a consequence of living and operating in the data-first world and delved deeper into each to understand it and suggest practical ways to address it. Let me briefly touch upon the essence of each paradox.

Mastering ten paradoxes to maximize value from data



The overarching '**data deluge vs insights drought**' paradox epitomizes the challenge of having vast data resources but not

having relevant insights for informed decision-making. It emphasizes the importance of extracting actionable insights from data rather than mere accumulation of data. A number of principles stated below are aimed at resolving this paradox, so that the data becomes a source of transformational value and not frustration. The '**infrastructure vs problem-solving**' paradox highlights the need to strike a balance between investing in scalable data infrastructure and applying logical problem-solving techniques. Both are crucial for realizing the full potential of data and extracting value from it, hence finding the right balance is key.

Then there is the '**short-term vs long-term**' paradox that highlights the challenge of building long-term capabilities while addressing the immediate business requirements. This emphasizes the need to adopt a two-speed approach, to balance quick, short-term problem-solving with long-term capability building, to build the necessary agility to operate in this dynamic, fast changing world. The '**open-source vs proprietary**' data paradox that underscore the importance of leveraging vast and open sources of data to generate comprehensive and deep insights, at the same time creating proprietary knowledge that would help build differentiation and competitive advantage. '**Data democratization and security**' paradox emphasizes the need to balance widespread data access critical for growth and collaboration, with robust security measures to keep the data safe, which is one of the biggest challenges in the digital age.

Moving on, the '**gut instinct vs data-driven decision-making**' paradox revolves around the tension between making decisions based on intuition and experience (gut), an age-old way of living and working, versus relying on data and analytics for more objective choices, highlighting the challenge of finding the right balance between these two approaches in decision-making. The '**hyper-personalization vs scaling**' paradox highlights the historical struggle of having to choose between individuality and scale, either ask for personalization that wasn't possible at scale, or live with standardized, common-for-all products and offerings. With AI, it is now possible to achieve personalization at scale.

The '**digital engagement vs digital detox**' reveals that the very digital engagement that brings multiple benefits to us, can also lead to burnout, stress and a sense of being overwhelmed, if it becomes excessive and constant, necessitating digital detox. Another one is '**sharing vs competitive advantage**', which underscores the importance of data for both collaborative growth, as well as establishing competitive advantage, creating a perpetual dilemma on whether to freely share data for progress or protect proprietary data to maintain a competitive edge. And last but not least, the long ongoing '**man vs machine**' debate over whether AI and automation will render humans irrelevant or serve as complementary tools to enhance human capabilities, highlighting the need to shift to a more collaborative narrative of AI enhancing human capabilities, that would require a shift from core data skills to problem-solving, business domain expertise and storytelling.

Dealing with these paradoxes is about finding your balance and making the choices within a given context. There will never be one solution that fits all. Throughout the book I have highlighted and brought out these paradoxes, I also want you to understand that the ultimate goal is to find your own balance. The key is to recognize that it is rarely about either/or but about **AND**. It is not that you have to choose one objective over the other, you can progress on both. And this is really what mastering paradoxes is about. To do that, it is important to get to the underlying assumptions behind each of the paradoxes and critically examine them. This brings clarity on the unifying objectives, which is fundamental to finding the right balance. And to achieve that, one has to go inside, anchor oneself.

Like in life, paradoxes exist in data, and they will always exist. The aim should be to master them, and not be limited by them. And I believe there are a few fundamental principles which when incorporated in our thinking, and actions, can help unlock the transformational value from data. Following are the **ten principles** to help break the paradoxes and provide practical guidelines for your data journey in the AI age:

- 1. Treat data as a source of value, not just input:**
Through this book, I hope I have succeeded in changing your perspective about data. You must now be able to rethink how you look at data in your organization and your lives. It's not merely an input or something collected and put in a system, a process, but a dynamic source of value. Data is your value generator across the entire DIAI framework, starting from data itself, the insights derived from it, the actions taken based on those insights and the ultimate impact. It is therefore critical that you consciously shift your mindset accordingly and treat data as your most valuable asset.
- 2. Start with the outcome in mind:** To realize transformational value from data, one needs to start with the end in mind. One needs to clearly define the outcomes expected. By setting a clear destination, you establish a purpose and direction for your data-related efforts. This razor-sharp focus on outcomes serves as a guiding compass, ensuring that you remain on course and do not get lost in the endless labyrinth of the data deluge.
- 3. Adopt product thinking:** Data is fluid in nature, which necessitates structuring it into products to effectively minimize the value leakage across the various stages of the data-to-insights-to-action cycle. Product thinking is to have an end-to-end focus and to integrate various elements across the data stack to deliver value in a repeatable manner. And given the omnipresence of data in the AI age, I believe that all products will need to incorporate elements across technology, domain, data and AI.

Ten principles to unlock the transformational value of data

- | | | | |
|----|---|----|--|
| 01 | Treat data as a source of value, not just input | 06 | Democratize access to data to unleash its power |
| 02 | Start with the outcome in mind to drive value from data | 07 | Balance data democratization and data security/privacy |
| 03 | Adopt a product thinking, the key to drive end to end focus | 08 | Leverage AI to deliver hyper-personalization at scale |
| 04 | Take an agile two-speed approach for short term focus while building long term capabilities | 09 | Collaborate to win by solving complex problems |
| 05 | Be AI first in solving for any data problem you are dealing with | 10 | Seek to gain competitive advantage with data |

4. **Take agile two-speed approach:** For effective execution of data initiatives, organizations must adopt a two-speed approach, which enables organizations to respond to short-term needs through Speed One and build long-term Speed Two capabilities by connecting the Speed One initiatives.
5. **Be AI-first:** AI and data have a recursive relationship. This has been the key theme across all the thirteen elements of the unified solutions framework I introduced in Section II. High-quality data enables transformational AI. And AI powered by this high-quality data can help solve the multiple challenges across the data management lifecycle; for example, data quality itself and particularly data security.
6. **Democratize access to data:** By making the right data available to the right people, at the right time, one can unleash the power of data for all. Democratization of access to information is a fundamental characteristic of the digital age, because it levels the playing field by providing widespread and equitable access to information, enabling better decision-making at all levels in the organization. Break the silos to unleash the full potential of data.

- 7. Balance data democratization with security/privacy:** Data privacy and security are emerging as the biggest threats—across enterprise, individuals and macro levels—making it harder to realize the opportunities for collaboration and growth that are enabled by data democratization and sharing. Just like in life, it is not one size fits all. The key is finding the right balance based on the context that would help streamline and tailor the security measures effectively.
- 8. Leverage AI to deliver hyper-personalization at scale:** We are now in the era of individualized experiences, segment of one, where every interaction can be tailored to the unique preferences and requirements of each person, be it customers or employees. This is a paradigm shift. Utilize AI to gain a deep understanding of these individual needs and deliver personalized solutions.
- 9. Collaborate to win:** Be it at the individual level, enterprise level or at a macro level, collaboration is critical. Data sharing allows access to information that individuals or organizations may not be able to generate independently, thus enhancing the depth and quality of the available data. This can help solve complex problems and drive innovation, which no one individual, enterprise or nation can drive in isolation. We are in an interdependent world, collaboration is the path to winning for all!
- 10. Seek to gain competitive advantage with data:** Leverage the vast potential of data for breakthrough innovation and transformation at all levels—individual, enterprise, nations. We are in the midst of the data revolution and entering an explosive and rapid phase of change with the AI age. Data, which is the foundation to it all, will be a source of competitive advantage. It is essential for everyone to adopt a strategic perspective,

an open mindset, a commitment to learning, and a focus on implementation to fully grasp this opportunity.

Some of these principles may sound like life lessons, which they are. Data is not just an enabler of the AI age, it mirrors life in both its beauty and challenges, and of course its paradoxes. Let us now explore what life lessons we can learn from the exploration of the data-first world we have gone through.

3. Are there some principles that we can learn from life and apply to data and vice versa, as we navigate the new world of data-enriched lives?

I began the chapter by saying life is data and data is life. Let me explain how.

I believe that life and data are intertwined or closely related. Most people are unaware of this profound interconnectedness of life and data. Even before the advent of digital, the core essence of our being has always been an intricate manifestation of data—our DNA—a biological data-storage system of encoded genetic blueprints that shape an organism's characteristics and functions. Our inherent cognitive process involves collection and storage of data by transforming our sensory inputs into memories and experiences. Conversely, data plays a pivotal role in helping us understand, preserve and enhance life. Be it our personal data, data on the environment, like temperature, humidity, etc., data on our history, health data, location data, financial data and the list goes on.

Data has always been a part of nature. From the classic golden ratio to the guiding mechanism of migratory birds, there is data in nature that defines us as humans. The leaf of a tree, a human ear, the design of a peacock's feather, the sting of a scorpion, all are in a golden ratio of 1:1.61 or phi. Similarly, the Vitruvian man as analysed by Leonardo da Vinci is a Fibonacci series—a progression of the golden ratio.* Nature and the planet we live on, are not random (well, we may have been formed as a result of the Big Bang, but

there is a method to this madness), it is full of data and math. So, it is fair to say that the principles of data we have explored in this book are in fact the principles of life.

Five principles for life and data



To me, there are certain principles of life that one applies to managing data paradoxes and vice versa. Life and data merge as both of these are journeys to reach a destination. What applies to one, applies to the other. They are no different from each other, just the context, scale and impact will differ. Some of the principles we can learn from life and apply to data are:

Do not avoid a technology wave whose time has come, make it your friend

Today, digital technologies have created a layer where data is visible, accessible and is in formats that we can manipulate. AI is adding to these capabilities. Data has always improved our lives, and it will continue to do so. And the first life lesson is **acceptance of the**

reality that is in front of us. Do not avoid technology and data. This wave has just come, **ride the wave.** In the given context, it is better that everyone opens themselves up to data, **make it your friend**, to be able to understand it, how to benefit from it, how to contribute to this revolution and how to improve it. This cycle has started, and we are that lucky generation that is living through the early stages of it. It is crucial that we develop capabilities, capacities and mindsets to open ourselves to the new wave that is here, and not hide in denial. Constant change has been the path of human civilization, and it will continue to be so!

Master the paradoxes by finding the right balance

Duality or paradoxes is the inherent nature of data and of life. Mastering multiple paradoxes is key to realizing transformational value from data. Same holds true for life. Do not take the paradoxes as a given. **To master the paradoxes, you have to find the right balance.** Treat paradoxes as an opportunity to delve deeper into a problem and to understand its essence. As you delve deeper, you are able to see the unifying objectives and how you can reframe the problem to break the paradoxes and achieve seemingly contradictory objectives. Those who master the paradoxes will break the boundaries and win!

Start with an end in mind but be open to change

We saw earlier that to derive transformational value from data initiatives, it is important that you start with the outcome in mind. That will ensure that you do not get lost in the data deluge and your efforts are focused. Similarly in life, unless you know the destination, you will keep wondering and perhaps also wandering. A sense of where you are going is important. You may change your path along the way, but you have to **start with the end in mind.** It will help you get moving with a sense of purpose, and not stay rooted in indecision or 'analysis-paralysis'. And once you move, new paths might emerge. And it is important that you keep yourself open to

learning and exploring those new paths, all the while staying focused on your destination. Another paradox!

Less is more

The promise of the Big Data world is about near-infinite scale. However, to make progress, you often have to narrow the problem first and 'thin slice' the data. Similarly in life, **less is often more**. Life is often about simplification and shedding things—minimalism! The wiser you get, the importance of possessions diminishes, the understanding deepens. 'Less is more', is often a core principle of life taught in many spiritual traditions. The work done by the South African scientists to find the beta variant—with little data—exemplified, how less is significant. The innovation that comes out of frugality may be more effective. Do not get greedy to collect more and more, learn to focus on and appreciate what is right in front of you.

As is the microcosm so is the macrocosm, as is the macrocosm so is the microcosm

I believe in the fractal nature of life and principles that align across levels—from individual, to enterprise, to society and beyond. The context will vary, the scale will vary, but principles remain the same. Not only can this concept be a great source of insight and foresight, when these levels start working in synchronicity, the impact is transformational. The task for us is to find the patterns and parallels. Whether in the context of data or life, we find this link between macro and micro layers. In the book, I detail the individual, macro and enterprise layers, and give you a glimpse of the patterns. I leave it to you to imagine how and where all can you apply this principle of '*yatha pinde tatha brahmande*' ('As is the individual, so is the Universe. As is the Universe, so is the individual.') from the ancient Indian text of Yajurveda.

As I think of data and life, these are the five principles I would like each one of us to go back with.

Carpe diem!

As I complete the journey of this book, I am filled with an incredible sense of excitement. In the eighteen months that it took me to write this book, the world around us has changed in so many amazing ways. Even eighteen months ago, the data-first world was already a reality. However, during this period, the AI age has truly come into its own with the explosion of Gen AI. The possibilities going forward are just mind-boggling. The troika of digital-data-AI is unstoppable and their impact on all aspects of life and business is just going to accelerate.

Do not sit on the sidelines when such momentous changes are happening in the world. Embrace the data-first world and the AI age. Yes, there are many challenges and uncertainties, and of course the many paradoxes to resolve. But do not let them bog you down. These paradoxes provide us with opportunities to push our thinking, to innovate, to transform and to create a better future at all levels—individual, enterprise and the world. So, make data your friend and strive boldly into the AI age!

Acknowledgements

The journey of writing *Mastering the Data Paradox* (MDP) has been rather different from my first book, *Winning in the Digital Age* (WITDA). Digital transformation, the focus of WITDA, has been in play for a while and is a topic I have been writing and speaking on for almost ten years. Whereas, data and AI, the subjects of MDP, are fast-evolving areas about which a lot is still unexplored. So, while writing WITDA was a lot about consolidating my past writings and experiences, MDP has been an extensive research-driven effort for the past two to three years. Here, I would like to acknowledge the core team members of the Incedo Insight Institute, Rahul Kumar, Ida Nair Sharma, Vikram T., Surbhi Mehta and Kanika Singal who have been an integral part of this book journey. This team has worked tirelessly by my side conducting in-depth research on every topic and ensuring a thorough and comprehensive exploration of the subject matter. Your seamless collaboration as a team and your collective commitment ensured that we created a work of depth and high quality.

I would also like to acknowledge the pivotal role played by many colleagues in the Incedo leadership team, who shared their industry and client expertise that provided the real-world perspective on the topics discussed. Thank you, Ashish Gupta, Manit Seth, Nihal Baghchandani, Yatin Bhatia and Archie Jackson for lending your expertise that helped me in adding depth to many topics in this book.

Beyond Incedo, I would like to thank my dear friend Dr Anurag Agrawal, dean, biosciences and health research, Ashoka University for your invaluable insights and practical examples on criticality of global data collaboration. Gratitude also to Dr Surya Tahora, professor at S.P. Jain Institute of Management and Research for your

profound insights on the topic of wisdom and leadership critical to various segments in Section III of the book.

I would also like to extend my gratitude to my clients at Incedo for allowing us to be the strategy to execution partners in their data transformation journey. The learnings from these hands-on experiences have helped in shaping and enriching my perspective on the concepts that I have talked about in this book.

I also extend my gratitude to my former colleagues at McKinsey, Fidelity, Flipkart and ActiveKarma for the invaluable lessons learned during our shared experiences that have shaped my understanding of the digital and data world and, indeed, life itself.

A special mention to Incedo's marketing and creative team, led by Rajat Shrimal, for helping me with all the commercial aspects of the book. And finally, a big thank you to the editors, Radhika Marwah and Saba Nehal, from Penguin Random House India for your editorial insights and willingness to speak your mind which made you an invaluable partner in this endeavour.

GLOSSARY OF DATA CONCEPTS



Scan the QR code to be directed to the glossary on the official website of *Mastering the Data Paradox*

* Data explosion estimates: It is not easy to measure the exact increase in data volume over the years, especially since there weren't many studies done before International Data Corporation (IDC) started tracking data systematically. So, I have used multiple estimates to measure data growth over time for comparison. Back in 2000, researchers at UC Berkeley, Hal R. Varian and Peter Lyman, made the first attempt at quantifying the amount of digital information in the world. In their study, 'How much information,' they provided two sets of estimates. If we consider the lower estimates, which include data adjusted for duplication and compression, the data has expanded from 0.00058 ZB in 1999 to 120 ZB in 2023, registering around 200,000 times growth in the past twenty-four years. If we consider the upper estimates, which include all the raw digital data, data has grown from 0.0017 ZB to 120 ZB, exhibiting about 70,000-fold increase since 1999. In this book, we used an approximate range of 1,00,000 to 1,50,000, because it's not possible to pinpoint the exact number.

- * Apache Hadoop is an open-source software platform that manages data processing and storage for Big Data applications. It enables distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to multiple machines with local computation and storage capacity.
- * MongoDB is a non-relational document database developed by MongoDB Inc. Launched in February 2009, MongoDB is used by web applications and other applications that need to store and query large amounts of data. It has a flexible data model that enables storage of unstructured data and provides full indexing support and replication with rich and intuitive APIs.
- † Apache Kafka is a tool that enables real-time data streaming and scalable data integration. It is widely used in various industries for building data pipelines, event streaming and building real-time data processing applications. Apache Kafka is an open-source solution for collection, processing and analysis of real-time, streaming data. It is designed to handle data streams from multiple sources and deliver them to multiple consumers.
- ‡ Snowflake is a cloud-based data platform that helps organizations store, manage and analyse large amounts of data efficiently. It is designed to handle structured and semi-structured data and provides a flexible and scalable solution for data storage and processing. It eliminates the need for upfront hardware investment and offers on-demand scalability.
- * Databricks is a cloud-based platform that simplifies and accelerates data analytics and machine learning tasks. It provides a collaborative environment for data professionals to work together on data-driven projects. Databricks offers powerful tools for data processing, data exploration and building machine learning models. It enables users to efficiently process and analyse large volumes of data.

- * Amazon Kinesis is a managed service provided by AWS that enables collection, processing and analysis of real-time, streaming data. Kinesis is designed to ingest, process and analyse streams of data in real time. As it is a managed service, it requires minimal setup and configuration efforts by the customer. It is cloud-based and hence scalable.
- † Apache Spark is an open-source unified analytics engine for large-scale data processing, including applications for batch processing, streaming and machine learning. Spark works by breaking down data into smaller pieces that can be processed in parallel, allowing faster speed. Spark also supports a wide variety of data formats, hence it can be used for data from a variety of sources.
- * In the context of machine learning, 'features' are pieces of data that are used to train and evaluate machine-learning models. A feature store is a repository for storing and managing pre-computed, reusable and high-quality features which helps to improve the efficiency and accuracy of ML models.

- * Apache Flink is an open-source, unified stream-processing and batch-processing framework developed by the Apache Software Foundation. It is used to process data streams at a large scale and to deliver real-time analytical insights from processed data.
- † ClickHouse is an open-source, column-oriented database-management system for online analytical processing that allows users to generate analytical reports using SQL queries in real-time. It is used for real-time analytics, real-time business reporting, data warehousing and Big Data analytics.
- ‡ Tinybird is a real-time data platform that helps businesses to make sense of their data in real-time by processing and analysing large amounts of data quickly and efficiently. The platform enables users to ingest, transform and analyse data from various sources, such as databases, APIs and streaming platforms.
- § Tecton is an enterprise-ready feature store platform to facilitate machine-learning operations. The platform is designed to streamline the process of creating, sharing and serving machine learning features, making it easier for data science and engineering teams to develop and deploy machine-learning models at scale.
- * Google BigQuery is a serverless data warehouse. It enables users to focus on data analytics without managing servers or storage, as Google manages the underlying infrastructure. Google BigQuery's popularity has been growing rapidly since its launch in 2011.
- † Azure Synapse or Azure Synapse Analytics is a cloud-based analytics service provided by Microsoft. It brings together Big Data and data warehousing capabilities into a single, unified platform. The platform provides tools and functionalities for data ingestion, data preparation, data warehousing and data integration. It enables users to easily integrate data from various sources, perform advanced analytics and build machine-learning models.
- ‡ Amazon Redshift is AWS's fully managed, petabyte-scale data-warehouse service in the cloud. It is designed to handle large volumes of data and enable fast and efficient data analysis. It uses

the same SQL-based tools and BI applications that are used on typical data warehouses.

- * IBM Db2 Warehouse is an industry leading data warehouse platform. Originally released in 2001 as IBM Netezza, it was designed as a high-performance, scalable data warehouse based on a massively parallel processing (MPP) architecture. In 2017, IBM released a cloud-based version of Db2 Warehouse that could be deployed and managed without the need for any on-premises infrastructure.
- † Oracle Autonomous Data Warehouse or (Oracle ADW) is a fully managed cloud-database service first released in 2019. It automates administration of a data warehouse, allowing users to completely focus on data applications.
- * TensorFlow is a free and open-source software library developed by the Google Brain team for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. TensorFlow Distributed is its distributed computing platform.

- * Pega is a software platform that helps businesses automate and streamline their processes and workflows through a set of tools and capabilities to design, build and deploy applications that can manage and optimize various business activities. Pega helps companies to create applications that handle tasks such as customer service, case management, workflow automation and business process management. It enables businesses to automate processes, improve productivity and make smarter decisions based on data.
- † Appian is a software platform that enables businesses to build and deploy custom applications quickly and easily. It helps businesses to design and automate various business processes, manage data and integrate with other systems. Appian offers a range of pre-built components and templates that can be customized to suit specific business needs. It allows for the creation of applications for tasks like project management, customer relationship management (CRM) and workflow automation. Appian enables businesses to automate processes, manage data and improve efficiency in various areas of their operations.

- * Microsoft Power BI is a business intelligence tool that helps organizations analyse and visualize their data in a user-friendly way. It allows users to connect to various data sources, such as databases, spreadsheets and cloud services, and transform that data into interactive reports and dashboards. Power BI also provides collaboration and sharing capabilities.
- † Google Data Studio (now Looker Studio) is one of Google's business intelligence (BI) tools. It provides web-based data visualization and facilitates customized dashboards and easy-to-understand reports.
- ‡ Trifacta is a software for data exploration and self-service data preparation on cloud and on-premises data platforms. It transforms and enriches raw data into clean and structured formats utilizing techniques in machine learning, data visualization, human-computer interaction and parallel processing.
- § Paxata is a self-service data preparation software that gets data ready for analytics. It combines data from different sources and checks for data-quality issues, such as duplicates and outliers. Algorithms and machine learning are used to automate certain aspects of data preparation, which is made available to users through a user-interface like Excel spreadsheets.
- * Qlik Sense is a cloud-based data-visualization product that can connect to a variety of data sources to create flexible and interactive visualizations.
- † Looker is a cloud-based business intelligence (BI) platform that can connect to a variety of data sources to create interactive dashboards and reports.

* An Art of Living silence retreat is a four–five-day residential retreat, where one is encouraged to disconnect with the external world, hence no phones are allowed; meditation and simple living is encouraged.

^{*} CERN or Conseil Européen pour la Recherche Nucléaire is French for the European Council for Nuclear Research.

^{*} As this book goes to publication, we bear witness to what is widely regarded as one of the most significant lapses in Israeli intelligence, failing to anticipate the attack by Hamas. Even the most sophisticated intelligence network has proven to be insufficient.

* The golden ratio is 1.618, represented by the Greek letter 'phi', which is said to be a mathematical connection between two aspects of an object. It is also called the Fibonacci sequence and it can be found across all of nature—in plants, animals, weather structures, star systems—it is ever-present in the universe.

Notes

Introduction: AI Age and the Data-First World

- 1 Zia Hayat, 'Digital trust: How to unleash the trillion-dollar opportunity for our global economy', WEF (17 August 2022), accessed on 14 July 2023, viewable at:
<https://www.weforum.org/agenda/2022/08/digital-trust-how-to-unleash-the-trillion-dollar-opportunity-for-our-global-economy/>.
- 2 'Largest American companies by market capitalization', Companies Market Cap, accessed on 14 July 2023, viewable at:
<https://companiesmarketcap.com/usa/largest-companies-in-the-usa-by-market-cap/>.
- 3 'What is the maximum number of columns / variables in statistics?', IBM Support (16 April 2020), accessed on 14 July 2023, viewable at: <https://www.ibm.com/support/pages/what-maximum-number-columns-variables-statistics/>.
- 4 Sajjad Hossan, 'The history and rise of Flipkart: Largest ecommerce company in India', Business Inspection (27 May 2023), accessed on 14 July 2023, viewable at:
<https://businessinspection.com.bd/history-and-rise-of-flipkart/>.
- 5 Klaus Schwab, 'The fourth Industrial Revolution: what it means, how to respond', WEF (14 January 2016), accessed on 14 July 2023, viewable at:
<https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>.

Section I: Understanding the Data-First World

Chapter 1: Data Explosion

- 1 Petroc Taylor, 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025', Statista (June 2021), accessed on 6 July 2023, viewable at:
<https://www.statista.com/statistics/871513/worldwide-data-created/>.
- 2 Julia Faria, 'Most valuable brands worldwide in 2023', Statista (January 2023), accessed on 6 July 2023, viewable at:
<https://www.statista.com/statistics/264875/brand-value-of-the-25-most-valuable-brands/>.
- 3 Peter Lyman and Hal R. Varian, 'How much information', School of Information Management and Systems at the University of California at Berkeley (Year 2000), accessed on 6 November 2023, viewable at: <https://groups.ischool.berkeley.edu/archive/how-much-info/how-much-info.pdf>.
- 4 'World population by year', Worldometer, accessed on 17 August 2023, viewable at: <https://www.worldometers.info/world-population/world-population-by-year/>.
- 5 Aaron O'Neill, 'Global gross domestic product (GDP) at current prices from 1985 to 2028', Statista (10 May 2023), accessed on 17 August 2023, viewable at:
<https://www.statista.com/statistics/268750/global-gross-domestic-product-gdp/>.
- 6 Petroc Taylor, 'Global mobile subscriptions since 1993', Statista (February 2023), accessed on 7 November 2023, viewable at:
<https://www.statista.com/statistics/262950/global-mobile-subscriptions-since-1993/>; 'Ericsson mobility report update: Global 5G subscriptions close to 1.3 billion in Q2 2023', Digital News Asia (2 September 2023), accessed on 7 November 2023, viewable at:
<https://www.digitalnewsasia.com/digital-economy/ericsson-mobility-report-update-global-5g-subscriptions-close-13-billion-q2-2023>.

- 7 'Internet growth statistics', Internet World Stats, accessed on 17 August 2023, viewable at:
<https://www.internetworldstats.com/emarketing.htm>; Ani Petrosyan, 'Digital population worldwide', Statista (October 2023), accessed on 7 November 2023, viewable at:
<https://www.statista.com/statistics/617136/digital-population-worldwide/>.
- 8 '10 Real World Applications of Internet of Things (IoT) – Explained in Videos', AnalyticsVidhya (20 June 2023), accessed on 19 December 2023, viewable at:
<https://www.analyticsvidhya.com/blog/2016/08/10-youtube-videos-explaining-the-real-world-applications-of-internet-of-things-iot/>; Lionel Sujay Vailshery, Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2023, with forecasts from 2022 to 2030', Statista (July 2023), accessed on 17 August 2023, viewable at:
<https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/>; Satyajit Sinha, 'State of IoT 2023: Number of connected IoT devices growing 16% to 16.7 billion globally', IoT Analytics (24 May 2023), accessed on 12 December 2023, viewable at: <https://iot-analytics.com/number-connected-iot-devices/>.
- 9 'How much information', School of Information Management and Systems, University of California at Berkeley, (Year 2000), accessed on 6 November 2023, viewable at:
<https://groups.ischool.berkeley.edu/archive/how-much-info/how-much-info.pdf>; Lucas Mearian, 'A zettabyte by 2010: Corporate data grows fiftyfold in three years', Computer world (6 March 2007), accessed on 17 August 2023, viewable at:
<https://www.computerworld.com/article/2543787/a-zettabyte-by-2010--corporate-data-grows-fiftyfold-in-three-years.html>; Petroc Taylor, 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025', Statista (June 2021), accessed on 7 November 2023, viewable at:

<https://www.statista.com/statistics/871513/worldwide-data-created/>.

- [10](#) George Firican, 'The history of big data', Lightsodata, accessed on 7 November 2023, viewable at:
<https://www.lightsodata.com/the-history-of-big-data/>.
- [11](#) 'How Big Data was used by the Romans for properties', Enriched Data (8 October 2020), accessed on 17 July 2023, viewable at:
<https://www.enricheddata.com/blog/how-big-data-was-used-by-the-romans-for-properties/>.
- [12](#) 'Why was the Roman army so powerful?', History Skills, accessed on 17 July 2023, viewable at:
<https://www.historyskills.com/classroom/year-7/year-7-roman-army-reading/>.
- [13](#) Rein Taagepera, 'Size and duration of empires: Growth-decline curves, 600 B.C. to 600 A.D.' *Social Science History* 3, no. 3/4 (1979): 115–38. <https://doi.org/10.2307/1170959>.
- [14](#) Alfredo Morabia, 'Epidemiology's 350th Anniversary: 1662–2012', National Library of Medicine (May 2013), accessed on 6 July 2023, viewable at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3640843/>.
- [15](#) 'Herman Hollerith', National Inventors Hall of Fame, accessed on 6 July 2023, viewable at:
<https://www.invent.org/inductees/herman-hollerith>.
- [16](#) Paul A. Freiberger, Michael R. Swaine, Erik Gregersen, William L. Hosch, Gloria Lotha, Shiveta Singh and Amy Tikkanen, 'ENIAC computer', Britannica, accessed on 17 July 2023, viewable at:
<https://www.britannica.com/technology/ENIAC>.
- [17](#) David Taylor, 'What is data warehouse? Types, definition & example', Guru99 (10 June 2023), accessed on 6 July 2023, viewable at: <https://www.guru99.com/data-warehousing.html#2>.
- [18](#) 'How LA used Big Data to build a Smart City in the 1970s', Architexturez (23 June 2015), accessed on 17 July 2023, viewable at: <https://architexturez.net/pst/az-cf-169297-1435054977>.
- [19](#) Tracey Wallace, 'How Amazon and independent ecommerce brands grew online sales 18233% in 20 Years', Big Commerce, accessed on 6 July 2023, viewable at:

<https://www.bigcommerce.com/blog/amazon-timeline-infographic/>.

- 20 Felix Richter, 'Amazon Maintains Lead in the Cloud Market', Statista (8 August 2023), accessed on 15 December 2023, viewable at: <https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/>.
- 21 'Google turns 20: how an internet search engine reshaped the world', Verge (27 September 2018), accessed on 15 December 2023, viewable at:
<https://www.theverge.com/2018/9/5/17823490/google-20th-birthday-anniversary-history-milestones>.
- 22 Petroc Taylor, 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025', Statista (June 2021), accessed on 18 August 2023, viewable at:
<https://www.statista.com/statistics/871513/worldwide-data-created/>.
- 23 Michael O'Reilly, 'The unseen data conundrum', *Forbes* (3 February 2022), accessed on 31 July 2023, viewable at:
<https://www.forbes.com/sites/forbestechcouncil/2022/02/03/the-unseen-data-conundrum/>.
- 24 George Michael, 'The Univac 1 Computer', Computer History, accessed on 17 August 2023, viewable at: <https://www.computer-history.info/Page4.dir/pages/Univac.dir/#:~:text=Remember%2C%20the%20UNIVAC%20was%20a,bits%20for%20decimal%20number%20representation>.
- 25 'Amazing facts and figures about the evolution of hard disk drives', Solarwinds Pingdom (10 July 2019), accessed on 17 August 2023, viewable at:
[https://www.pingdom.com/blog/amazing-facts-and-figures-about-the-evolution-of-hard-disk-drives/#:~:text=IBM%20introduced%20the%20first%20hard,550%20pounds%20\(250%20kg\).](https://www.pingdom.com/blog/amazing-facts-and-figures-about-the-evolution-of-hard-disk-drives/#:~:text=IBM%20introduced%20the%20first%20hard,550%20pounds%20(250%20kg).)
- 26 Luca Clissa, 'How big Is Big Data in 2021?' (2 March 2023), Medium, accessed on 17 August 2023, viewable at:
<https://towardsdatascience.com/how-big-are-big-data-in-2021->

[6dc09aff5ced](#); Sébastien Stormacq, 'Celebrate Amazon S3's 17th birthday at AWS Pi Day 2023', AWS News (14 March 2023), accessed on 17 August 2023, viewable at:

<https://aws.amazon.com/blogs/aws/celebrate-amazon-s3s-17th-birthday-at-aws-pi-day-2023/>.

[27](#) John Rydning, 'Global DataSphere', IDC, accessed on 6 July 2023, viewable at: https://www.idc.com/getdoc.jsp?containerId=IDC_P38353.

[28](#) Ogi Djuraskovic, 'Big Data Statistics 2023: How Much Data is in The World?', firstsiteguide.com (4 October 2023), accessed on 20 December 2023, viewable at: <https://firstsiteguide.com/big-data-stats/>; Branka Vuleta. 'How Much Data Is Created Every Day? +27 Staggering Stats', (28 October 2021), accessed on 20 December 2023, viewable at: <https://seedscientific.com/how-much-data-is-created-every-day/>; Petroc Taylor, 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025', Statista (June 2021), accessed on 15 December 2023 viewable at:
<https://www.statista.com/statistics/871513/worldwide-data-created/>.

[29](#) Nikolina Cveticanin, '30 Big Data statistics everybody's talking about', Data Prot, (updated on 14 July 2023), accessed on 6 July 2023, viewable at: <https://dataprot.net/statistics/data-statistics/>.

[30](#) Maryam Mohsin, '10 Google Search Statistics', Oberlo (13 January 2023), accessed on 17 August 2023, viewable at:
<https://www.oberlo.com/blog/google-search-statistics>.

[31](#) Liz Alton, 'How video is reshaping digital advertising', X Business (formerly Twitter), accessed on 17 August 2023, viewable at:
<https://business.twitter.com/en/blog/how-video-is-reshaping-digital-advertising.html>.

[32](#) Branka, 'Netflix statistics – 2023', Truelist (9 January 2023), accessed on 17 August 2023, viewable at:
<https://truelist.co/blog/netflix-statistics/#:~:text=Netflix%20viewership%20statistics%20show%20that,solely%20statistics%20for%20US%20users>; Prateek Saxena, '51+ Netflix Statistics & Facts Worth Knowing',

Appinventiv (1 June 2022), accessed on 21 December 2023, viewable at: <https://appinventiv.com/blog/netflix-statistics-facts/#:~:text=Netflix%20has%202221.64%20Million%20Paid%20Subscribers%20Worldwide>.

- 33 Stacy Jo Dixon, 'Most popular social networks worldwide as of January 2023, ranked by number of monthly active users', Statista (July 2023), accessed on 17 August 2023, viewable at: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>.
- 34 Jimit Bagadiya, '200+ Social Media Statistics and facts of 2023', SocialPilot (6 July 2023), accessed on 17 August 2023, viewable at: <https://www.socialpilot.co/blog/social-media-statistics>.
- 35 Aditya Rayaprolu, 'How much data is created every day in 2023?', Techjury (26 July 2023), accessed on 17 August 2023, viewable at: <https://techjury.net/blog/how-much-data-is-created-every-day/>.
- 36 Petroc Taylor, 'Data volume of Internet of Things (IoT) connections worldwide in 2019 and 2025', Statista (September 2022), accessed on 21 July 2023, viewable at: <https://www.statista.com/statistics/1017863/worldwide-iot-connected-devices-data-size/>.
- 37 Statista Research Department, 'Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions)', Statista (November 2016), accessed on 21 July 2023, viewable at: <https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/>.
- 38 Fire Industry Association, 'A history of the Internet of Things', Fire Industry Association (6 October 2021), accessed on 6 July 2023, viewable at: <https://www.fia.uk.com/news/history-of-iot.html>.
- 39 Zia Hayat, 'Digital trust: How to unleash the trillion-dollar opportunity for our global economy', WEF (17 August 2022), accessed on 6 July 2023, viewable at: <https://www.weforum.org/agenda/2022/08/digital-trust-how-to-unleash-the-trillion-dollar-opportunity-for-our-global-economy/>; Arya Devi, 'DCO 2030: Digital economy to contribute 30% of global GDP and create 30 million jobs by 2030', Edge Middle East

(5 February 2023), accessed on 6 July 2023 viewable at:
<https://www.edgemiddleeast.com/business/dco-2030-digital-economy-to-contribute-30-of-global-gdp-and-create-30-million-jobs-by-2030>.

40 Gil Press, 'A very short history of big data', *Forbes* (9 May 2013), accessed on 17 August 2023, viewable at:
<https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/>; Fabio Duarte, 'Amount of data created daily (2023)', Exploding Topics (3 April 2023), accessed on 17 August 2023, viewable at: <https://explodingtopics.com/blog/data-generated-per-day>.

41 Gil Press, 'WhatsApp statistics, users, demographics as of 2023', What's the Big Data (5 December 2023), accessed on 20 December 2023, viewable at:
<https://whatsthebigdata.com/whatsapp-statistics/>.

42 Jimit Bagadiya, '200+ social media statistics and facts of 2023', Social Pilot (7 November 2023), accessed on 20 December 2023, viewable at: <https://www.socialpilot.co/blog/social-media-statistics>.

43 Shubham B. Rajput, 'Big Data 3 V's and 5 V's', Medium (24 June 2020), accessed on 6 July 2023, viewable at:
<https://medium.com/analytics-vidhya/big-data-3-vs-and-5-v-s-c1cae2a6d311>.

44 Terry Brown, 'Who is using Big Data in business?', IT Chronicles (12 March 2021), accessed on 6 July 2023, viewable at:
<https://itchronicles.com/big-data/who-is-using-big-data-in-business/>.

45 '57 Amazon statistics to know in 2023', Landing Cube (20 December 2022), accessed on 6 July 2023, viewable at:
<https://landingcube.com/amazon-statistics/>.

46 'The Digitization of the World, from Edge to Core', Seagate (November 2018), accessed on 6 July 2023, downloadable from:
<https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>.

47 'Research: 71% of tech leaders confirm a clear link between use of real-time data and revenue growth', Businesswire (17 August

2022), accessed on 6 July 2023, viewable at:

<https://www.businesswire.com/news/home/20220817005204/en/Research-71-of-Tech-Leaders-Confirm-A-Clear-Link-Between-Use-of-Real-Time-Data-and-Revenue-Growth>.

48 Bernard Marr, 'What's the difference between structured, semi-structured and unstructured data?', *Forbes* (18 October 2019), accessed on 17 August 2023, viewable at:

<https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/>.

49 'Data growth is real, and 3 other key findings', Matillion (26 January 2022), accessed on 6 July 2023, viewable at:

<https://www.matillion.com/resources/blog/matillion-and-idg-survey-data-growth-is-real-and-3-other-key-findings/>.

Chapter 2: Data, the Fuel for the Digital Age

1 Janice Fernandes, 'How important is in-car GPS to global consumers?', YouGov (10 August 2022), accessed on 2 November 2023, viewable at: <https://business.yougov.com/content/43393-how-important-car-gps-global-consumers>.

2 Natasha Lomas, 'Israel passes emergency law to use mobile data for COVID-19 contact tracing', TechCrunch (18 March 2020), accessed on 17 August 2023, viewable at:
<https://techcrunch.com/2020/03/18/israel-passes-emergency-law-to-use-mobile-data-for-covid-19-contact-tracing/>.

3 '23 Binge-watching statistics you should know (2023)', *Soocial.com*, accessed on 17 August 2023, viewable at:
<https://www.soocial.com/binge-watching-statistics/>.

4 'Global Over-The-Top (OTT) market anticipated to garner a revenue of \$438.5 billion by 2026 and rise at a CAGR of 19.1% over the analysis timeframe 2019 to 2026', GlobalNewswire (6 September 2022), accessed on 14 July 2023, viewable at:
<https://www.globenewswire.com/en/news-release/2022/09/06/2510521/0/en/Global-Over-The-Top-OTT-Market-Anticipated-to-Garner-a-Revenue-of-438-5-Billion-by-2026->

[and-Rise-at-a-CAGR-of-19.1-over-the-Analysis-Timeframe-2019-to-2026-150-Pages-research-Dive.html](#).

- 5 'Global Smart Wearable market - market to grow by 19.48% from 2021 - 2026', Business Wire (8 February 2021), accessed on 17 August 2023, viewable at:
<https://www.businesswire.com/news/home/20210208005342/en/Global-Smart-Wearable-Market---Market-to-Grow-by-19.48-from-2021---2026--->
ResearchAndMarkets.com#:~:text=According%20to%20Cisco%20Systems%2C%20the,that%20suits%20the%20everyday%20lifestyle.
- 6 Oleg Bestsennyy, Greg Gilbert, Alex Harris and Jennifer Rost, 'Telehealth: A quarter-trillion-dollar post-COVID-19 reality?', McKinsey and Company (9 July 2021), accessed on 14 July 2023, viewable at: <https://www.mckinsey.com/industries/healthcare/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality/>.
- 7 Simon Kemp, 'Facebook users, stats, data & trends', Datareportal (11 May 2023), accessed on 14 July 2023, viewable at: <https://datareportal.com/essential-facebook-stats/>.
- 8 'Countries in the world by population', Worldometer, accessed on 17 August 2023, viewable at: <https://www.worldometers.info/world-population/population-by-country/>.
- 9 '23% of users in U.S. say social media led them to change views on an issue; some cite Black Lives Matter', Pew Research (15 October 2020), accessed on 14 July 2023, viewable at: <https://www.pewresearch.org/fact-tank/2020/10/15/23-of-users-in-us-say-social-media-led-them-to-change-views-on-issue-some-cite-black-lives-matter/>.
- 10 Julie Thompson, 'What marketers need to know about people's social media patterns during the pandemic', Business (22 February 2023), accessed on 14 July 2023, viewable at: <https://www.business.com/articles/social-media-patterns-during-the-pandemic/>.
- 11 Ethan Cramer-Flood, 'Worldwide Ecommerce forecast update 2022', Insider Intelligence (29 July 2022), accessed on 17 August

2023, viewable at:

<https://www.insiderintelligence.com/content/worldwide-ecommerce-forecast-update-2022>.

- 12 'Impact of COVID pandemic on eCommerce', International Trade Administration, accessed on 14 July 2023, viewable at:
<https://www.trade.gov/impact-covid-pandemic-ecommerce>.
- 13 'COVID-19 drives global surge in use of digital payments', World Bank (29 June 2022), accessed on 14 July 2023, viewable at:
<https://www.worldbank.org/en/news/press-release/2022/06/29/covid-19-drives-global-surge-in-use-of-digital-payments>.
- 14 'COVID-19 set to radically accelerate digital transformation in the retail banking industry', BCG (6 May 2020), accessed on 14 July 2023, viewable at: <https://www.bcg.com/press/6may2020-covid-set-to-accelerate-digital-transformation-in-retail-banking-industry>.
- 15 Statista Research Department, 'Number of active online banking users worldwide in 2020 with forecasts from 2021 to 2024, by region', Statista (March 2021), accessed on 14 July 2023, viewable at: <https://www.statista.com/statistics/1228757/online-banking-users-worldwide/>.
- 16 'How the digital surge will reshape finance', *Economist* (8 October 2020), accessed on 14 July 2023, viewable at:
<https://www.economist.com/finance-and-economics/2020/10/08/how-the-digital-surge-will-reshape-finance>.
- 17 Peter High, 'Former Amazon exec Marco Argenti drives a remarkable digital transformation at Goldman Sachs', *Forbes* (25 January 2023), accessed on 14 July 2023, viewable at:
<https://www.forbes.com/sites/peterhigh/2023/01/25/former-amazon-exec-marco-argenti-drives-a-remarkable-digital-transformation-at-goldman-sachs/>.
- 18 'FinTech lending market', Allied Market Research (October 2021), accessed on 14 July 2023, downloadable from:
<https://www.alliedmarketresearch.com/fintech-lending-market-A14263>.
- 19 'The global fraud detection and prevention market forecast', Fortune Business Insights (May 2023), accessed on 14 July 2023,

downloadable from:

<https://www.fortunebusinessinsights.com/industry-reports/fraud-detection-and-prevention-market-100231>.

20 'Digital payment market to hit opportunities worth \$361.30 billion by 2030: Grand View Research, Inc.', Bloomberg (19 September 2022), accessed on 14 July 2023, viewable at:

<https://www.bloomberg.com/press-releases/2022-09-19/digital-payment-market-to-hit-opportunities-worth-361-30-billion-by-2030-grand-view-research-inc/>.

21 'Market insights - Shared mobility users – Worldwide', Statista (October 2023), accessed on 14 July 2023, viewable at:

<https://www.statista.com/outlook/mmo/shared-mobility/worldwide#users>.

22 'Mobility as a Service market', Markets and Markets (August 2023), accessed on 14 July 2023, downloadable from:

<https://www.marketsandmarkets.com/Market-Reports/mobility-as-a-service-market-78519888.html>.

23 Stephanie Chevalier, 'Projected retail e-commerce GMV share of Amazon in the United States from 2016 to 2021', Statista (April 2017), accessed on 14 July 2023, viewable at:

<https://www.statista.com/statistics/788109/amazon-retail-market-share-usa/>.

24 Don Davis, 'Deep discounts drive online Black Friday sales up 7.5%', Digitalcommerce360 (25 November 2023), accessed on 4 January 2024, viewable at:

<https://www.digitalcommerce360.com/article/black-friday-ecommerce-sales/>.

25 Anna Baluch, '38 E-commerce statistics of 2023', *Forbes* (8 February 2023), accessed on 14 July 2023, viewable at:

https://www.forbes.com/advisor/business/ecommerce-statistics/#sources_section.

26 Sameer Chhabra, 'Netflix says 80 percent of watched content is based on algorithmic recommendations', Mobil Syrup (22 August 2017), accessed on 17 August 2023, viewable at:

<https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/>.

- 27 Brad Adgate, 'As Media Companies Focus On Streaming, The Audience Of Their Cable Networks Continue To Drop', *Forbes* (5 January 2022), accessed on 20 December 2023, viewable at: <https://www.forbes.com/sites/bradadgate/2022/01/05/as-media-conglomerates-focus-on-streaming-the-audience-of-their-cable-networks-continue-to-drop/>.
- 28 Brad Adgate, 'Global ad revenue for print struggles, as total ad revenue nears \$1 trillion', *Forbes* (7 March 2023), accessed on 14 July 2023, viewable at: <https://www.forbes.com/sites/bradadgate/2023/03/07/global-ad-revenue-for-print-struggles-as-total-ad-revenue-nears-1-trillion/>.
- 29 J. G. Navarro, 'Distribution of global advertising expenditure from 2015 to 2022, by media', Statista (July 2022), accessed on 14 July 2023, viewable at: <https://www.statista.com/statistics/245440/distributuion-of-global-advertising-expenditure-by-media/>.
- 30 J. G. Navarro, 'Distribution of global advertising expenditure from 2015 to 2022, by media', Statista (July 2022), accessed on 14 July 2023, viewable at: [https://www.statista.com/statistics/245440/distributuion-of-global-advertising-expenditure-by-media/.\)](https://www.statista.com/statistics/245440/distributuion-of-global-advertising-expenditure-by-media/.))
- 31 'Personalized medicine market outlook', Future Market Insights (January 2023), accessed on 14 July 2023, viewable at: <https://www.futuremarketinsights.com/reports/personalized-medicine-market>.
- 32 Paul J. Lee, Mayank NK Choudhary and Ting Wang, 'Online resources for studies of genome biology and epigenetics', *Current Opinion in Toxicology* 6, October 2017, pp 34–41 accessed on 20 December 2023, viewable at: <https://www.sciencedirect.com/science/article/abs/pii/S2468202017300621>.
- 33 'BASF's lab assistant, designed to simplify your lab life', BASF, accessed on 14 July 2023, viewable at: <https://dispersions-resins.bASF.com/emea/en/lab-assistant-to-make-your-formulation-easier-than-ever-.html>.

Chapter 3: Value Reimagined

- 1 World Wildlife Fund, 'Water scarcity overview', accessed on 7 July 2023, viewable at: <https://www.worldwildlife.org/threats/water-scarcity>.
- 2 'The Data Age is here, Are You Ready?', ESG Research Insight Paper, Splunk, accessed on 4 January 2023, viewable at: https://www.splunk.com/en_us/pdfs/gated/ebooks/data-age.pdf.
- 3 Tom Kevan, 'Aerospace Digital Twins gain altitude', Digital Engineering Magazine, accessed on 10 July 2023, viewable at: https://bt.e-ditionsbyfry.com/publication/frame.php?i=738580&p=&pn=&ver=html5&view=articleBrowser&article_id=4218180.
- 4 'Everything You Need to Know about UPS Route Optimization Software (ORION)', Route4me (20 May 2021), accessed on 22 August 2023, viewable at: <https://blog.route4me.com/ups-route-optimization-software-orion/>.
- 5 Bernard Marr, 'The brilliant ways UPS uses artificial intelligence, machine learning and Big Data', *Forbes* (15 June 2018), accessed on 22 August 2023, viewable at: <https://www.forbes.com/sites/bernardmarr/2018/06/15/the-brilliant-ways-ups-uses-artificial-intelligence-machine-learning-and-big-data/>.
- 6 'The next era of essential intelligence', Annual Report 2021, S&P Global (25 February 2022), accessed on 10 July 2023, downloadable from: https://s29.q4cdn.com/690959130/files/doc_downloads/annual-shareholder-meeting/2021/S-P-Global-2021-Annual-Report.pdf.
- 7 Tiago Bianchi, 'Advertising revenue of Google from 2001 to 2022', Statista (10 September 2023), accessed on 19 December 2023, viewable at: <https://www.statista.com/statistics/266249/advertising-revenue-of-google/>.
- 8 Jitender Miglani, 'Amazon sales and profit analysis for 2022: top 10 insights', Forrester (21 February 2023), accessed on 22 August

2023, viewable at: <https://www.forrester.com/blogs/amazon-sales-and-profit-analysis-for-2022-top-10-insights/>.

9 Mansoor Iqbal, 'Uber revenue and usage statistics (2023)', Business for Apps (20 February 2023), accessed on 10 July 2023, viewable at: <https://www.businessofapps.com/data/uber-statistics/>.

10 Mathilde Carlier, 'Tesla's revenue from FY 2008 to FY 2022', Statista (January 2023), Accessed on 10 July 2023, viewable at: <https://www.statista.com/statistics/272120/revenue-of-tesla/>.

11 James Tyrrell, 'Synthetic data for AI fills gaps in edge cases', TechHQ (18 October 2022), accessed on 22 August 2023, viewable at: <https://techhq.com/2022/10/synthetic-data-for-ai-fills-gaps-in-autonomous-vehicle-edge-cases/>.

12 Tom Davenport and Maryam Alavi, 'How to train generative AI using your company's data', HBR (6 July 2023), accessed on 22 August 2023, viewable at: <https://hbr.org/2023/07/how-to-train-generative-ai-using-your-companys-data>.

13 Rosane Giovis and Eniko Rozsa, 'Transforming customer service: How generative AI is changing the game', IBM (17 July 2023), accessed on 22 August 2023, viewable at: [https://www.ibm.com/blog/transforming-customer-service-how-generative-ai-is-changing-the-game/](https://www.ibm.com/blog/transforming-customer-service-how-generative-ai-is-changing-the-game).

14 Michael Chui et al, 'The economic potential of generative AI: The next productivity frontier', McKinsey (14 June 2023), accessed on 18 August 2023, viewable at: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights>.

15 Tom Davenport and Maryam Alavi, 'How to train generative AI using your company's data', HBR (6 July 2023), accessed on 22 August 2023, viewable at: <https://hbr.org/2023/07/how-to-train-generative-ai-using-your-companys-data>.

16 'How Big Data Analysis helped increase Walmarts Sales turnover?', ProjectPro (27 October 2023), accessed on 4 January 2024, viewable at: <https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109>; Meredith

Flora, 'Walmart Supply Chain: What Makes It (Still) So Successful', ShipBob (19 August 2022), accessed on 4 January 2023, viewable at: <https://www.shipbob.com/blog/walmart-supply-chain/>.

Chapter 4: The Data Paradox

- 1 'The Data Paradox - Research Findings', Dell (2020), accessed on 7 July 2023, downloadable from: <https://www.dell.com/en-us/dt/perspectives/data-paradox.htm#footnote-ref1&pdf-overlay=/www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/industry-market/data-paradox-research-findings.pdf>.
- 2 Sarah Brinks, 'Unveiling data challenges afflicting businesses around the world', Forrester Consulting (May 2021), accessed on 7 July 2023, downloadable from: <https://www.delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/data-paradox-forrester-thought-leadership-paper.pdf>.
- 3 'How much data is created every day and how to collect it', CIO bulletin (4 April 2022), accessed on 21 July 2023, viewable at: <https://www.ciobulletin.com/big-data/how-much-data-is-created-every-day-and-how-to-collect-it>.
- 4 'The Data Paradox - Research Findings', Dell (2020), accessed on 7 July 2023, downloadable from: <https://www.dell.com/en-us/dt/perspectives/data-paradox.htm#footnote-ref1&pdf-overlay=/www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/industry-market/data-paradox-research-findings.pdf>.
- 5 R. Buckminster Fuller, Kiyoshi Kuromiya and Adjuvant, *Critical Path*, (New York: Macmillan, 1982).
- 6 Paul H. Silvergate and David Jarvis, 'Data: a double-edged sword', Deloitte Insights (15 September 2022), accessed on 2 November 2023, available at: <https://www2.deloitte.com/us/en/insights/industry/technology/challenges-in-data-management.html>; Sarah Brinks, 'Unveiling Data challenges afflicting businesses around the world', Forrester and

Dell Technologies (May 2021), accessed on 10 July 2023, downloadable from: <https://www.delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/data-paradox-forrester-thought-leadership-paper.pdf>.

- 7 Petroc Taylor, 'Big data market size revenue forecast worldwide from 2011 to 2027', Statista (March 2018), accessed on 22 August 2023, viewable at:
<https://www.statista.com/statistics/254266/global-big-data-market-forecast/>.
- 8 'State of Trust Report 2023', Vanta, (accessed on 8 November 2023, viewable at: <https://8588479.fs1.hubspotusercontent-na1.net/hubfs/8588479/2023%20State%20of%20Trust%20Report.pdf>).
- 9 Petroc Taylor, 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025', Statista (June 2021), accessed on 22 August 2023, viewable at:
<https://www.statista.com/statistics/871513/worldwide-data-created/>.
- 10 Laurence van der Sande, 'Why you need to capitalize on the rise of the data-driven enterprise', Accenture (21 May 2021), accessed on 10 July 2023, viewable at: <https://www.accenture.com/nl-en/blogs/insights/data-driven-enterprise>.
- 11 'Why do 87% of data science projects never make it into production?', Venture Beat (19 July 2019), accessed on 22 August 2023, viewable at: <https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/>.
- 12 Brian T. O'Neill, 'Failure rates for analytics, AI, and big data projects = 85% – yikes!', Designing for Analytics (23 July 2019), accessed on 10 July 2023, viewable at:
<https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/>.
- 13 Nitin Jayaraman, 'Unifying business data: breaking down data silos', Cloud IO (12 May 2020), accessed on 22 August 2023, viewable at: <https://www.cloudio.io/unifying-business-data-breaking-down-data-silos/>.

- 14 Thomas H. Davenport, et al, 'Analytics and AI-driven enterprises thrive in the age of with: The culture catalyst', Deloitte (25 July 2019), accessed on 10 July 2023, viewable at:
<https://www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html>.
- 15 'The Data Paradox - Research Findings', Dell (2020), accessed on 10 July 2023, downloadable from: <https://www.dell.com/en-us/dt/perspectives/data-paradox.htm#footnote-ref1&pdf-overlay=/www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/industry-market/data-paradox-research-findings.pdf>.
- 16 '2020 State of SaaS Ops', BetterCloud, accessed on 10 July 2023, viewable at: https://pages.bettercloud.com/rs/719-KZY-706/images/2020_StateofSaaSOpsReport.pdf.
- 17 Michael Shirer and Jessica Goepfert, 'Global spending on big data and analytics solutions will reach \$215.7 billion in 2021, according to a new IDC spending guide', Business Wire (17 August 2021), accessed on 12 July 2023, viewable at:
<https://www.businesswire.com/news/home/20210817005182/en/Global-Spending-on-Big-Data-and-Analytics-Solutions-Will-Reach-215.7-Billion-in-2021-According-to-a-New-IDC-Spending-Guide1>.
- 18 Brian T. O'Neill, 'Failure rates for analytics, AI, and big data projects = 85% – yikes!', Designing for Analytics (23 July 2019), accessed on 12 July 2023, viewable at:
<https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/>.
- 19 *Digital Transformation*, Raconteur (30 September 2021), accessed on 12 July 2023, downloadable from:
<https://res.cloudinary.com/yumyoshjin/image/upload/v1632924841/pdf/digital-transformation-2021.pdf>.
- 20 Paramita (Guha) Ghosh, 'Prescriptive analytics use cases by Paramita Ghosh', Dataversity (27 April 2021), accessed on 12 July 2023, viewable at: <https://www.dataversity.net/prescriptive-analytics-use-cases/>.
- 21 'Breaking down the data language barrier', Sigma Computing (6 January 2021), accessed on 12 July 2023, viewable at:

<https://www.sigmacomputing.com/blog/breaking-down-the-data-language-barrier>.

- 22 'Empower your overwhelmed data team: IT roundtable', GDS Group (19 July 2022), accessed on 12 July 2023, viewable at: <https://gdsgroup.com/events/roundtable/empower-your-overwhelmed-data-team/>.
- 23 Susan Moore, 'Gartner Data shows 87 percent of organizations have low BI and analytics maturity', Gartner (6 December 2018), accessed on 12 July 2023, viewable at: <https://www.gartner.com/en/newsroom/press-releases/2018-12-06-gartner-data-shows-87-percent-of-organizations-have-low-bi-and-analytics-maturity>.
- 24 Maria Korolov, 'Unlocking the hidden value of dark data', CIO (11 August 2022), accessed on 12 July 2023, viewable at: <https://www.cio.com/article/404526/unlocking-the-hidden-value-of-dark-data.html>.
- 25 Leandro DalleMule and Thomas H. Davenport, 'What's your data strategy?', HBR (May–June 2017), accessed on 12 July 2023, viewable at: <https://hbr.org/2017/05/whats-your-data-strategy>.
- 26 Sarah Brinks, 'Unveiling Data challenges afflicting businesses around the world', Forrester and Dell technologies (May 2021), accessed on 10 July 2023, downloadable from: <https://www.dell.com/en-us/dt/perspectives/data-paradox.htm#pdf-overlay=/www.delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/data-paradox-forrester-thought-leadership-paper.pdf>.

Chapter 5: The Root Cause

- 1 'The cost of not modernizing your infrastructure monitoring', Logic Monitor (1 September 2022), accessed on 12 July 2023, viewable at: <https://www.logicmonitor.com/blog/the-cost-of-not-modernizing-your-infrastructure-monitoring>.
- 2 Insight Direct, 'The state of IT modernization 2020', IDG (2020), accessed on 12 July 2023, downloadable from:

<https://solutions.insight.com/getattachment/a67b34bd-1a9a-42fe-a408-7afe180b96d8/Complete-IDG-survey-results.aspx>.

- 3 Manasi Sakpal, 'How to improve your data quality', Gartner (14 July 2021), accessed on 12 July 2023, viewable at:
<https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality/>.
- 4 Brian T. O'Neill, 'Failure rates for analytics, AI, and big data projects = 85% – yikes!', Designing for Analytics (23 July 2019), accessed on 12 July 2023, viewable at:
<https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/>.
- 5 Susan Moore, 'Gartner Data shows 87 percent of organizations have low BI and analytics maturity', Gartner (6 December 2018), accessed on 12 July 2023, viewable at:
<https://www.gartner.com/en/newsroom/press-releases/2018-12-06-gartner-data-shows-87-percent-of-organizations-have-low-bi-and-analytics-maturity>.
- 6 Lindsay McGuire, 'Digital Transformation: are data silos ruining your productivity?', Formstack (1 July 2020), accessed on 12 July 2023, viewable at: <https://www.formstack.com/blog/how-data-silos-hurt-productivity>.
- 7 'Gartner Survey: CDOs are delivering business impact and enabling digital transformation', ITnation (6 December 2017), accessed on 12 July 2023, viewable at: <https://itnation.lu/news/gartner-survey-cdos-are-delivering-business-impact-and-enabling-digital-transformation/>.
- 8 Chris Brock, 'What's the ceiling on the adoption of business intelligence?', Pyramid Analytics (14 June 2022), accessed on 12 July 2023, viewable at:
<https://www.pyramidanalytics.com/blog/whats-the-ceiling-on-the-adoption-of-business-intelligence/>.

Section II: Maximizing Value in the Data-First World

Chapter 7: Define the Business Problems

- 1 Charles Conn, Hugo Sarrazin and Simon London, 'How to master the seven-step problem-solving process', McKinsey (13 September 2019), accessed on 14 July 2023, viewable at:
<https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-to-master-the-seven-step-problem-solving-process>.
- 2 Kelly Blum, Matt LoDolce and Gloria Omale, 'Gartner marketing survey finds only 14% of organizations have achieved a 360-degree view of their customer', Gartner (19 January 2022), accessed on 14 July 2023, viewable at:
<https://www.gartner.com/en/newsroom/press-releases/gartner-marketing-survey-finds-only-14--of-organizations-have-ac>.

Chapter 8: Multi-Source Data

- 1 Sara Brown, 'Why external data should be part of your data strategy', MIT Sloan (18 February 2021), accessed on 14 July 2023, viewable at: <https://mitsloan.mit.edu/ideas-made-to-matter/why-external-data-should-be-part-your-data-strategy>.
- 2 'Samsung: A next-level mobile payment experience', Solaris SE, accessed on 14 July 2023, viewable at:
<https://www.solarisgroup.com/en/case-studies/samsung/>.
- 3 Douglas McCarthy, 'State all-payer claims databases: tools for improving health care value, part 1', Commonwealth Fund (10 December 2020), accessed on 14 July 2023, viewable at:
<https://www.commonwealthfund.org/publications/fund-reports/2020/dec/state-apcds-part-1-establish-make-functional>;
Stephanie Cohen and Lynn Quincy, 'All-payer claims databases: Unlocking data to improve healthcare value', Altarum Healthcare Value Hub (September 2015), accessed on 14 July 2023 and viewable at: <https://www.healthcarevaluehub.org/advocate-resources/publications/all-payer-claims-databases-unlocking-data-improve-health-care-value>.

- 4 'Fortune 500 Companies are Now Relying on Social Media for Their Research', Unbox Social, accessed on 19 December 2023, viewable at: <https://www.unboxsocial.com/blog/fotune-500-companies-research/#:~:text=Almost%2099%25%20of%20the%20Fortune,to%20understand%20their%20target%20market>.
- 5 Louis Columbus, 'The state of enterprise data integration, 2020', *Forbes* (29 March 2020), accessed on 14 July 2023, viewable at: <https://www.forbes.com/sites/louiscolumbus/2020/03/29/the-state-of-enterprise-data-integration-2020/>.
- 6 'Data growth is real, and 3 other key findings', Matillion and IDG (26 January 2022), accessed on 14 July 2023, viewable at: <https://www.matillion.com/resources/blog/matillion-and-idg-survey-data-growth-is-real-and-3-other-key-findings>.

Chapter 9: Real-Time Data

- 1 Louie Andre, '53 important statistics about how much data is created every day', Finance Online (16 November 2023), accessed on 10 November 2023, viewable at: <https://financesonline.com/how-much-data-is-created-every-day/>.
- 2 Ani Petrosyan, 'Number of internet and social media users worldwide as of April 2023 (in billions)', Statista (July 2023), accessed on 14 July 2023, viewable at: <https://www.statista.com/statistics/617136/digital-population-worldwide/>.
- 3 David Reinsel, 'How you contribute to today's growing Datasphere and its enterprise impact', IDC (4 November 2019), accessed on 10 November 2023, viewable at: <https://blogs.idc.com/2019/11/04/how-you-contribute-to-todays-growing-datasphere-and-its-enterprise-impact/#:~:text=IDC%20expects%20there%20to%20be,smart%20home%20devices%20and%20wearables>.
- 4 'IoT growth demands rethink of long-term storage strategies, says IDC', IoTBusinessNews (29 July 2020), accessed on 12 September 2023, viewable at:

<https://iotbusinessnews.com/2020/07/29/20898-iot-growth-demands-rethink-of-long-term-storage-strategies-says-idc/>.

- 5 'What is streaming data?', AWS, accessed on 14 July 2023, viewable at: <https://aws.amazon.com/streaming-data/>.
- 6 Brad Brown and Josh Gottlieb, 'The need to lead in data and analytics', McKinsey (21 April 2016), accessed on 21 July 2023, viewable at: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-need-to-lead-in-data-and-analytics>.
- 7 Ian MacKenzie, Chris Meyer and Steve Noble, 'How retailers can keep up with consumers', McKinsey (1 October 2013), accessed on 17 July 2023, viewable at:
<https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers>.
- 8 Christopher Robinson, 'Guide to real-time customer service', Finance Online, accessed on 17 July 2023, viewable at:
<https://financesonline.com/guide-to-real-time-customer-service/>.
- 9 Mercer Smith, '107 customer service statistics and facts you shouldn't ignore', Help Scout (23 June 2023), accessed on 17 July 2023, viewable at: <https://www.helpscout.com/75-customer-service-facts-quotes-statistics/>.
- 10 Anubhav Bhattacharjee, Violet Chung, Shwaitang Singh and Renny Tho, 'Building a winning AI neobank', McKinsey (29 November 2022), accessed on 19 December 2023, viewable at:
<https://www.mckinsey.com/industries/financial-services/our-insights/building-a-winning-ai-neobank>.
- 11 'AWS supply chain features', AWS, accessed on 17 July 2023, viewable at: <https://aws.amazon.com/aws-supply-chain/features/>.
- 12 'Mobil ServSM real time', ExxonMobil Services, accessed on 17 July 2023, viewable at: <https://www.mobil.co.in/en-in/business/services/real-time>.
- 13 Jennifer Rainey Marquez, 'The COVID-19 Data plan: 3 innovative ways Johnson & Johnson is using data science to fight the pandemic', Johnson and Johnson (13 January 2021), accessed on 17 July 2023, viewable at: <https://www.jnj.com/innovation/how-johnson-johnson-uses-data-science-to-fight-covid-19-pandemic>.

- 14 'Information governance—challenges and solutions', OpenText (2023), accessed on 17 July 2023, viewable at:
https://www.microfocus.com/media/flyer/information_governance_challenges_and_solutions_flyer.pdf.
- 15 Bernard Marr, 'How to use real-time data? key examples and use cases', *Forbes* (14 March 2022), accessed on 14 July 2023, viewable at:
<https://www.forbes.com/sites/bernardmarr/2022/03/14/how-to-use-real-time-data-key-examples-and-use-cases/>.
- 16 Fabio Duarte, 'Amount of Data created daily (2023)', Exploding Topics (3 April 2023), accessed on 8 November 2023, viewable at:
<https://explodingtopics.com/blog/data-generated-per-day>.
- 17 'Real-time, real value: 80% of businesses see revenue increases thanks to real-time data', PR Newswire (9 May 2022), accessed on 17 July 2023, viewable at: <https://www.prnewswire.com/news-releases/real-time-real-value-80-of-businesses-see-revenue-increases-thanks-to-real-time-data-301542344.html>.
- 18 Inbal Aharoni, '87-percent of enterprises lack the budget to achieve game-changing analytics', Sqream (2 September 2021), accessed on 17 July 2023, viewable at:
<https://sqream.com/media-room/87-percent-of-enterprises-lack-the-budget-to-achieve-game-changing-analytics/>.

Chapter 10: Proprietary Data

- 1 Thomas H. Davenport and Thomas Redman, 'Getting advantage from proprietary data', *Wall Street Journal* (11 March 2015), accessed on 17 July 2023, viewable at:
<https://www.wsj.com/articles/BL-CIOB-6471>.
- 2 Diana Porumboiu, 'Accessing tacit knowledge to drive more innovation', Viima (18 March 2022), accessed on 15 September 2023, viewable at: <https://www.viima.com/blog/tacit-knowledge>.
- 3 Jeff Schwartz et al., 'Knowledge management: Creating context for a connected world', Deloitte (15 May 2020), accessed on 17 July 2023, viewable at:

<https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2020/knowledge-management-strategy.html>.

Chapter 11: Modern Data Stack

- 1 Sven Blumberg et al., 'Why you need a digital data architecture to build a sustainable digital business', McKinsey (13 November 2017), accessed on 17 July 2023, viewable at:
<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-you-need-a-digital-data-architecture>.
- 2 Databricks and Fivetran, 'Building a modern data stack of the future', accessed on 17 July 2023, downloadable from:
<https://get.fivetran.com/rs/353-UTB-444/images/building-a-modern-data-stack-of-the-future.pdf>.
- 3 Jeff Barr, 'The AWS Blog: The First Five Years', AWS News Blog (9 November 2009), accessed on 19 December 2023, downloadable from: <https://aws.amazon.com/blogs/aws/aws-blog-the-first-five-years/>; 'Announcing Amazon Redshift', AWS (28 November 2012), accessed on 19 December 2023, downloadable from:
<https://aws.amazon.com/about-aws/whats-new/2012/11/28/announcing-amazon-redshift>.
- 4 Joseph McKendrick, 'How Netflix built its real-time data infrastructure', Venture Beat (17 February 2022), accessed on 17 July 2023, viewable at: <https://venturebeat.com/data-infrastructure/how-netflix-built-its-real-time-data-infrastructure/>.
- 5 George Anadiotis, 'The move to modern data architecture: 2022 data delivery and consumption patterns survey', ChaosSearch (May 2022), accessed on 17 July 2023, downloadable from:
<https://www.chaossearch.io/hubfs/resources/DBTA%20Market%20Research%20Report%202022.pdf>.

Chapter 12: Data Quality

- 1 Zia Muhammad, 'Study shows google collects most data out of all big tech companies', Digital Information World (28 May 2022), accessed on 15 September 2023, viewable at:

<https://www.digitalinformationworld.com/2022/05/study-shows-google-collects-most-data.html>.

- 2 Daniel Shvartsman, 'Amazon: The world's most powerful economic and cultural force', [Investing.com](https://www.investing.com) (10 August 2023), accessed on 15 September 2023, viewable at:
<https://www.investing.com/academy/statistics/amazon-facts/>.
- 3 David Lazer, Ryan Kennedy, 'What we can learn from the epic failure of Google Flu Trends', *Wired* (1 October 2015), accessed on 17 July 2023, viewable at: <https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/>.
- 4 Alexis C. Madrigal, 'In defense of Google Flu Trends', *The Atlantic* (27 March 2014), accessed on 17 July 2023 and viewable at:
<https://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/>.
- 5 Ian MacKenzie, Chris Meyer and Steve Noble, 'How retailers can keep up with consumers', *McKinsey* (1 October 2013), accessed on 17 July 2023, viewable at:
<https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers>.
- 6 Michael Chui et al., 'The state of AI in 2022—and a half decade in review', *McKinsey* (6 December 2022), accessed on 17 July 2023, viewable at:
<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review>.
- 7 Meghan Rimol, 'Gartner survey reveals 80% of executives think automation can be applied to any business decision', *Gartner* (22 August 2022), accessed on 17 July 2023, viewable at:
<https://www.gartner.com/en/newsroom/press-releases/2022-08-22-gartner-survey-reveals-80-percent-of-executives-think-automation-can-be-applied-to-any-business-decision>.
- 8 Roger Magoulas and Steve Swoyer, 'The state of data quality in 2020', *O'Reilly* (12 February 2020), accessed on 17 July 2023, viewable at: <https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/>.
- 9 PayPal Editorial Staff, 'Adaptive Machine Learning: The future of ecommerce fraud management', *PayPal* (3 August 2022), Accessed

on 19 December 2023, accessible on:

<https://www.paypal.com/us/brc/article/ecommerce-fraud-management>.

10 Brian Eastwood, 'What is synthetic data—and how can it help you competitively?', MIT (23 January 2023), accessed on 17 July 2023, viewable at: <https://mitsloan.mit.edu/ideas-made-to-matter/what-synthetic-data-and-how-can-it-help-you-competitively>.

11 Rob Toews, 'Synthetic Data is about to transform artificial intelligence', *Forbes* (12 June 2022), accessed on 17 July 2023, viewable at:
<https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/>.

Chapter 13: Data Products

- 1 Tushar Haralkar, 'Getting ready for a new era of Data protection: what your business needs to know', IBM India Newsroom (17 October 2023), accessed on 28 November 2023, viewable at:
<https://in.newsroom.ibm.com/getting-ready-for-new-era-of-data-protection>.
- 2 Campbell Abbey and Nick Richmond, 'The Power of the Data-Driven Enterprise', Accenture (2019), accessed on 17 July 2023, viewable at: https://d1.awsstatic.com/executive-insights/en_US/ebook-accenture-the-power-of-the-data-driven-enterprise.pdf.
- 3 Maximilian Faschan, 'What if data became everybody's business?', Medium (9 January 2022), accessed on 28 October 2023, viewable at: <https://towardsdatascience.com/what-if-data-became-everybodys-business-85b7c20d6ab7>; 'You are wasting 90pc of your company's potential', Techwire Asia (3 July 2018), accessed on 28 October 2023, viewable at:
<https://techwireasia.com/2018/07/you-are-wasting-90-percent-of-your-companys-potential/#:~:text=EVERY%20day%2C%20the%20world%20produces,pile%2C%2090%20percent%20goes%20unused>.

- 4 'Unstructured Data', MongoDB, accessed on 28 October 2023, viewable at: <https://www.mongodb.com/unstructured-data>.
- 5 Leandro DalleMule and Thomas H. Davenport, 'What's your Data strategy?', *Harvard Business Review* (May–June 2017), accessed on 28 October 2023, viewable at: <https://hbr.org/2017/05/whats-your-data-strategy>.
- 6 Eric Siegel, 'Models are rarely deployed: An industry-wide failure in machine learning leadership', KD Nuggets (17 January 2022), accessed on 28 October 2023, viewable at: <https://www.kdnuggets.com/2022/01/models-rarely-deployed-industrywide-failure-machine-learning-leadership.html>.
- 7 Mark Tossell, '6 Reasons Why BI and Analytics Projects Fail – And How to Avoid It', SFBEN (22 November 2021), accessed on 28 October 2023, viewable at: <https://www.salesforceben.com/6-reasons-why-bi-and-analytics-projects-fail-and-how-to-avoid-it/>; Thomas Wood, '10 reasons why data science projects fail', Fast Data Science (18 May 2020), accessed on 28 October 2023, viewable at: <https://fastdatascience.com/why-do-data-science-projects-fail/>.
<https://fastdatascience.com/why-do-data-science-projects-fail/>
- 8 'Free dark data assessment', Konverge, accessed on 28 October 2023, viewable at: <https://www.konverge.com.au/dark-data>.
- 9 Simon O'Regan, 'Designing Data Products', Towards Data Science (16 August 2018), accessed on 16 September 2023, viewable at: <https://towardsdatascience.com/designing-data-products-b6b93edf3d23>.

Chapter 14: Agility

- 1 Randy Bean, 'Why is it so hard to become a data-driven company?', *Harvard Business Review* (5 February 2021), accessed on 20 September 2023, viewable at: <https://hbr.org/2021/02/why-is-it-so-hard-to-become-a-data-driven-company>.
- 2 Ibid.
- 3 Julia Prats et al., 'Organizational Agility', Oliver Wyman (April 2018), accessed on 17 July 2023, downloadable from:

https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2018/april/Organizational_Agility.pdf.

- 4 Simon Rogers, 'Catching up on 15 years of Google Trends', Google (11 August 2021), accessed on 17 July 2023, viewable at: <https://blog.google/products/search/catching-15-years-google-trends/>; Simon Rogers, 'What is Google Trends data—and what does it mean?', Medium (11 August 2021), accessed on 17 July 2023, viewable at: <https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8>.

Chapter 15: Data Democratization

- 1 Thomas H. Davenport et al., 'Analytics and AI-driven enterprises thrive in the Age of With', Deloitte (25 July 2019), accessed on 17 July 2023, viewable at: <https://www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html>.
- 2 Ani Petrosyan, 'Number of internet and social media users worldwide as of April 2023 (in billions)', Statista (July 2023), accessed on 17 July 2023, viewable at: <https://www.statista.com/statistics/617136/digital-population-worldwide/>.
- 3 MLW, 'AIRBNB: The growth story unveiled', Medium (1 June 2016), accessed on 25 September 2023, viewable at: <https://medium.com/dropbear-tech/airbnb-the-growth-story-unveiled-e40d7e2e9b3b>.
- 4 Matthew Woodward, 'Airbnb Statistics [2023]: user & market growth data', Search Logistics (18 July 2023), accessed on 25 September 2023, viewable at: <https://www.searchlogistics.com/learn/statistics/airbnb-statistics/>.
- 5 Chris C. Williams, 'Democratizing data at Airbnb', Medium (12 May 2017), accessed on 25 September 2023 viewable at: <https://medium.com/airbnb-engineering/democratizing-data-at-airbnb-852d76c51770>.
- 6 Keith D. Foote, 'A brief history of Business Intelligence', Dataversity (6 April 2023), accessed on 17 July 2023, viewable at:

<https://www.dataversity.net/brief-history-business-intelligence/>;

Cristina Lago, '150 years of business intelligence: A brief history' CIO (18 July 2018), accessed on 17 July 2023, viewable at: <https://www.cio.com/article/221963/history-of-business-intelligence.html>.

- 7 Laurence Goasdouff, 'Gartner Says Cloud Will Be the Centerpiece of New Digital Experiences', Gartner (10 November 2021), accessed on 19 December 2023, viewable at: <https://www.gartner.com/en/newsroom/press-releases/2021-11-10-gartner-says-cloud-will-be-the-centerpiece-of-new-digital-experiences>.
- 8 Ruhaab Markas, 'Ask Data: Simplifying analytics with natural language', Tableau (26 November 2018), accessed on 25 September 2023, viewable at: <https://www.tableau.com/blog/ask-data-simplifying-analytics-natural-language-98655>.
- 9 Emma Salomon 'Measuring the Productivity Impact of Generative AI', National Bureau of Economic Research (No. 6, June 2023), accessed on 19 December 2023, viewable at: <https://www.nber.org/digest/20236/measuring-productivity-impact-generative-ai>.

Chapter 16: Data Security

- 1 Siobhan Climer, 'History Of Cyber Attacks From The Morris Worm To Exactis', Mindsight (3 July 2018) accessed on 19 December 2023, viewable at: <https://gomindsight.com/insights/blog/history-of-cyber-attacks-2018/>.
- 2 Chaim Gartenberg, 'Security startup Verkada hack exposes 150,000 security cameras in Tesla factories, jails, and more', The Verge (10 March 2021), accessed on 27 September 2023, viewable at: <https://www.theverge.com/2021/3/9/22322122/verkada-hack-150000-security-cameras-tesla-factory-cloudflare-jails-hospitals>.
- 3 Jignasa Sinha, '5 artificial intelligence-based attacks that shocked the world in 2018', Analytics India Magazine (20 December 2018), accessed on 27 September 2023, viewable at:

<https://analyticsindiamag.com/5-artificial-intelligence-based-attacks-that-shocked-the-world-in-2018/>.

- 4 '2023 SonicWall Cyber Threat Report', SonicWall, accessed on 27 September 2023, viewable at: <https://www.sonicwall.com/2023-mid-year-cyber-threat-report/>.
- 5 'Cybercrime to cost the world 8 trillion annually in 2023', Cybercrime Magazine (17 October 2022), accessed on 14 November 2023, viewable at:
<https://cybersecurityventures.com/cybercrime-to-cost-the-world-8-trillion-annually-in-2023/>; Carmen Ene, '10.5 Trillion reasons why we need a united response to cyber risk', *Forbes* (22 February 2023), accessed on 27 September 2023, viewable at:
<https://www.forbes.com/sites/forbestechcouncil/2023/02/22/105-trillion-reasons-why-we-need-a-united-response-to-cyber-risk/>.
- 6 'The NHS cyber-attack', Acronis (7 February 2020), accessed on 27 September 2023, viewable at: <https://www.acronis.com/en-sg/blog/posts/nhs-cyber-attack/>.
- 7 Zaheer Merchant, 'NotPetya: the cyberattack that shook the world', *Economic Times* (5 March 2022), accessed on 27 September 2023, viewable at:
<https://m.economictimes.com/tech/newsletters/ettech-unwrapped/notpetya-the-cyberattack-that-shook-the-world/articleshow/89997076.cms>.
- 8 Rohit Chintapali, 'Satya Nadella vouches for zero-trust security approach', *Business World* (5 January 2023), accessed on 18 July 2023, viewable at: <https://www.businessworld.in/article/Satya-Nadella-Vouches-For-Zero-Trust-Security-Approach/05-01-2023-460625/>.
- 9 Aranza Trevino, Anne Cutler and Darren Guccione, 'How cybercriminals are using AI for cyberattacks', Keeper Security (21 June 2023), accessed on 27 September 2023, viewable at:
<https://www.keepersecurity.com/blog/2023/06/21/how-cybercriminals-are-using-ai-for-cyberattacks/>.
- 10 Jesse Damiani, 'A voice deepfake was used to scam a CEO out of \$243,000', *Forbes* (3 September 2019), accessed on 27 September 2023, viewable at:

<https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>.

- 11 'Cyber Security Industry Data', Fortuner Business Insights (April 2023), accessed on 27 September 2023, downloadable from: <https://www.fortunebusinessinsights.com/industry-reports/cyber-security-market-101165>.

Chapter 17: Organizational Alignment

- 1 Brian Gregg et al., 'The most perfect union: Unlocking the next wave of growth by unifying creativity and analytics', McKinsey (18 June 2018), accessed on 28 September 2023, viewable at: <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-most-perfect-union>.
- 2 'The warning signs and risks of poor organizational alignment', Lucid, accessed on 28 September 2023, viewable at: <https://lucid.co/blog/warning-signs-and-risks-of-poor-organizational-alignment>.
- 3 Thaisa Fernandes, 'Learn More About the Spotify Squad Framework—Part I', Medium (7 March 2017), accessed on 19 December 2023 and accessible at: <https://medium.com/pm101/spotify-squad-framework-part-i-8f74bcfc761>; Tim Brewer, 'How Spotify Organizes its Org Chart', Functionaly, accessed on 19 December 2023 and accessible at: <https://www.functionaly.com/originometry/real-org-charts/spotify-org-structure>; Michael Mankins and Eric Garton, 'How Spotify Balances Employee Autonomy and Accountability', HBR Business Management, (9 February 2017), Accessed on 19 December 2023 and available at: <https://hbr.org/2017/02/how-spotify-balances-employee-autonomy-and-accountability>.

Chapter 18: Data Culture

- 1 Thomas H. Davenport et al., 'Analytics and AI-driven enterprises thrive in the Age of With', Deloitte (25 July 2019), accessed on 29 September 2023, viewable at:

<https://www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html>.

- 2 Bernard Marr, 'Data-driven decision-making: beware of the HIPPO effect!', *Forbes* (26 October 2017), accessed on 18 July 2023, viewable at:
<https://www.forbes.com/sites/bernardmarr/2017/10/26/data-driven-decision-making-beware-of-the-hippo-effect/>.
- 3 Ibid.
- 4 Leo Nogas, 'Thinking beyond cultural legacy: the case of Korean Air', *Medium* (23 June 2013), accessed on 21 July 2023, viewable at: <https://leonogas.medium.com/thinking-beyond-cultural-legacy-the-case-of-korean-air-ac049b190b6d>; Travis Dixon, 'Power Distance and Plane Crashes: The Gladwell Hypothesis', *Thematic Education* (6 May 2020), accessed on 18 July 2013, viewable at: <https://www.thematic-education.com/ibpsych/2020/05/06/power-distance-and-plane-crashes-the-gladwell-hypothesis/>; Ji Yeon Song, 'The effects of cultural factors on safety in aviation focusing on asian and western cultures', Embry-Riddle Aeronautical University (April 2018), accessed on 18 July 2013, downloadable from:
<https://commons.erau.edu/cgi/viewcontent.cgi?article=1143&context=student-works>.
- 5 Tasha Eurich, 'Working with people who aren't self-aware', *Harvard Business Review* (19 October 2018), accessed on 15 November 2013, viewable at: <https://hbr.org/2018/10/working-with-people-who-arent-self-aware>.
- 6 Tasha Eurich, 'What self-awareness really is (and how to cultivate it)', *Harvard Business Review* (4 January 2018), accessed on 18 July 2013, viewable at: <https://hbr.org/2018/01/what-self-awareness-really-is-and-how-to-cultivate-it>.
- 7 Stuart R. Levine, 'Diversity confirmed to boost innovation and financial results', *Forbes* (15 January 2020), accessed on 18 July 2013, viewable at:
<https://www.forbes.com/sites/forbesinsights/2020/01/15/diversity-confirmed-to-boost-innovation-and-financial-results/>.

- 8 Saul Mcleod, 'Stanley milgram shock experiment: summary, results, & ethics', Simply Psychology (16 June 2023), accessed on 18 July 2013, viewable at:
<https://www.simplypsychology.org/milgram.html>.
- 9 'Resistance to change in organizations comes from these 5 factors', LeadershipIQ, accessed on 18 July 2013, viewable at:
<https://www.leadershipiq.com/blogs/leadershipiq/resistance-to-change-in-organizations-comes-from-these-5-factors-new-data/>.
- 10 'NewVantage Partners releases 2021 Big data and AI executive survey', Businesswire (4 January 2021), accessed on 18 July 2013, viewable at:
<https://www.businesswire.com/news/home/20210104005022/en/NewVantage-Partners-Releases-2021-Big-Data-and-AI-Executive-Survey/>.
- 11 Jamie Stober, 'How Airbnb is boosting data literacy with 'Data U Intensive' training', Medium (11 December 2018), accessed on 18 July 2023, viewable at: <https://medium.com/airbnb-engineering/how-airbnb-is-boosting-data-literacy-with-data-u-intensive-training-a6399dd741a2>.
- 12 Christopher Doering, 'How PepsiCo is harnessing data to help retailers increase sales', Food Dive (4 October 2021), accessed on 18 July 2023, viewable at: <https://www.fooddive.com/news/how-pepsi-co-is-harnessing-data-to-help-retailers-increase-sales/606464/>; Lisa Johnston, 'PepsiCo launching Pepviz data science practice for retail partners', Consumer Goods (9 August 2021), accessed on 18 July 2023 viewable at:
<https://consumergoods.com/pepsi-co-launching-pepviz-data-science-practice-retail-partners>.
- 13 Reza Reza, 'Uber's Big Data platform: 100+ petabytes with minute latency', Uber (17 October 2018), accessed on 18 July 2023, viewable at: <https://www.uber.com/en-IN/blog/uber-big-data-platform/>.
- 14 Ruth Umoh, 'Why Jeff Bezos makes Amazon execs read 6-page memos at the start of each meeting', CNBC (23 April 2018), accessed on 18 July 2023, viewable at:
<https://www.cnbc.com/2018/04/23/what-jeff-bezos-learned-from->

[requiring-6-page-memos-at-amazon.html](#); Jesse Freeman, 'The Anatomy of an Amazon 6-pager', Medium (16 July 2020), accessed on 18 July 2023, viewable at: <https://writingcooperative.com/the-anatomy-of-an-amazon-6-pager-fc79f31a41c9>.

15 Collin Bryar and Bill Carr, *Working Backwards: Insights, Stories, and Secrets from Inside Amazon* (New York: St. Martin's Press, 2021).

16 Jesse Freeman, 'The Anatomy of an Amazon 6-pager', Medium (16 July 2020), accessed on 18 July 2023, viewable at: <https://writingcooperative.com/the-anatomy-of-an-amazon-6-pager-fc79f31a41c9>.

17 Incedo, 'Incedo Belief System- Guiding principles that are core to everything we do', accessed on 18 July 2023, viewable at: <https://www.incedoinc.com/incedo-belief-system/>.

18 David Selinger, 'Data Driven: what Amazon's Jeff Bezos taught me about running a company', Entrepreneur (11 September 2014), accessed on 18 July 2023, viewable at: <https://www.entrepreneur.com/business-news/data-driven-what-amazons-jeff-bezos-taught-me-about/237326>.

19 Collin Bryar and Bill Carr, *Working Backwards: Insights, Stories, and Secrets from Inside Amazon* (New York: St. Martin's Press, 2021).

Chapter 19: Data Talent

1 'Big data talent shortage: How to bridge the gap?', Fractal, accessed on 18 July 2023, viewable at: <https://fractal.ai/news/big-data-talent-shortage-bridge-gap/>.

2 Thomas H. Davenport and DJ Patil, 'Is Data scientist still the sexiest job of the 21 century?', *Harvard Business Review* (15 July 2022), Accessed on 18 July 2023 and viewable at: <https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century>.

3 'Occupational Outlook Handbook: Data Scientists', U.S. Bureau of Labor Statistics (6 July 2023), accessed on 30 September 2023, viewable at: <https://www.bls.gov/ooh/math/data-scientists.htm>.

- 4 'The future of data science: Career outlook and industry trends', Tech Target (15 July 2022), accessed on 18 July 2023, viewable at: <https://www.techtarget.com/searchenterpriseai/feature/The-future-of-data-science-jobs>.
- 5 Nicolaus Henke, Jordan Levine and Paul McInerney, 'Analytics translator: The new must-have role', McKinsey (1 February 2018), accessed on 30 September 2023, viewable at: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/analytics-translator>.

Section III: Data for Individuals and Beyond

Chapter 21: The World of Hyper-Personalization

- 1 Richard Winger and David Edelman, 'Segment-of-One Marketing', BGC (1 January 1989), accessed on 19 July 2023, viewable at: <https://www.bcg.com/publications/1989/strategy-segment-of-one-marketing>.
- 2 Nidhi Arora et al., 'The value of getting personalization right—or wrong—is multiplying', McKinsey (12 November 2021), accessed on 2 August 2023, viewable at: <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying>.
- 3 Nicholas Lansing, 'The Next-Generation Wealth Manager', Forbes Insights, accessed on 19 July 2023, viewable at: <https://www.temenos.com/wp-content/uploads/2019/07/Forbes-report-19-Jun-19.pdf>.
- 4 '12 Hyper personalization statistics that demonstrate value', Monetate, accessed on 4 October 2023, viewable at: <https://monetate.com/resources/12-hyper-personalization-statistics-that-demonstrate-value/>.
- 5 Ann Binlot, 'Janett Nichol explains how Nike fuses together science and art to make innovative athletic apparel', *Forbes* (5 April 2019), accessed on 2 August 2023, viewable at: <https://www.forbes.com/sites/abinlot/2019/04/05/janett-nichol->

[explains-how-nike-fuses-together-science-and-art-to-make-innovative-athletic-apparel/](https://www.sciencedirect.com/science/article/pii/S089826601830313X).

- 6 'Become a lego designer', LEGO Ideas, accessed on 2 August 2023, viewable at: <https://ideas.lego.com/projects/create>.
- 7 'Why hyper-personalisation is the next big thing in banking', The Banker, accessed on 4 October 2023, viewable at: <https://www.thebanker.com/Why-hyper-personalisation-is-the-next-big-thing-in-banking-1680514156>.
- 8 'How AI hyper-personalization helps fintechs and financial services boost customer satisfaction', Techcrunch, accessed on 4 October 2023, viewable at: <https://techcrunch.com/sponsor/nvidia-aws-company/how-ai-hyper-personalization-helps-fintechs-and-financial-services-boost-customer-satisfaction/>.
- 9 Alex Barseghian, 'How Nike is using analytics to personalize their customer experience', *Forbes* (7 October 2019), accessed on 2 August 2023, viewable at: <https://www.forbes.com/sites/forbestechcouncil/2019/10/07/how-nike-is-using-analytics-to-personalize-their-customer-experience/>.
- 10 Tera Hatfield, 'Nike By You Studio', accessed on 2 August 2023, viewable at: <https://www.terahatfield.com/nike-studio>.
- 11 'NikeID history', Nice Kicks, accessed on 2 August 2023, viewable at: <https://www.nicekicks.com/nike-id/>.
- 12 Alex Attard-Manche, 'Nike's 3D Printing: Just Do It', Harvard: Technology and Operations Management (12 November 2018), Accessed on 2 August 2023, viewable at: <https://d3.harvard.edu/platform-rctom/submission/nikes-3d-printing-just-do-it/>.
- 13 'Customer Data Platform Market', Markets and Markets, accessed on 19 July 2023, viewable at: <https://www.marketsandmarkets.com/Market-Reports/customer-data-platform-market-94223554.html>.
- 14 Ken Robinson, 'Do schools kill creativity?', TED Talks (February 2006), accessed on 19 July 2023, viewable at: https://www.ted.com/talks/sir_ken_robinson_do_schools_kill_creativity.

- 15 'Guild', Britannica (30 September 2023), accessed on 2 August 2023, viewable at: <https://www.britannica.com/money/topic/guild-trade-association>.
- 16 Paul J. Lee, Mayank N.K. Choudhary, Ting Wang, 'Online resources for studies of genome biology and epigenetics', *Current Opinion in Toxicology* 6 (October 2017): 34–41, accessed on 17 November 2023, viewable at: <https://www.sciencedirect.com/science/article/abs/pii/S2468202017300621>.
- 17 'mRNA Vaccines and Therapeutics market size is projected to reach USD 58.92 billion by 2031, growing at a CAGR of 13.3%: Straits Research', Globe News Wire, accessed on 17 November 2023, viewable at: <https://www.globenewswire.com/news-release/2023/06/08/2684961/0/en/mRNA-Vaccines-and-Therapeutics-Market-Size-is-projected-to-reach-USD-58-92-billion-by-2031-growing-at-a-CAGR-of-13-3-Straits-Research.html>.

Chapter 22: Data for Better Decision-Making

- 1 Tim Stobierski, 'The Advantages of Data-Driven Decision-Making', *Harvard Business Review* (26 August 2019), accessed on 19 July 2023, viewable at: <https://online.hbs.edu/blog/post/data-driven-decision-making>; R. Kelly Garrett, 'Should we worry that half of Americans trust their gut to tell them what's true?', The Conversation (28 September 2017), accessed on 19 July 2023, viewable at: <https://theconversation.com/should-we-worry-that-half-of-americans-trust-their-gut-to-tell-them-whats-true-84259>.
- 2 'Intuition', Stanford Encyclopedia of Philosophy (4 December 2012), accessed on 21 July 2023 and accessible at: <https://plato.stanford.edu/entries/intuition/>.
- 3 Christine Ma-Kellams, Jennifer Lerner 'Trust Your Gut or Think Carefully? Examining Whether an Intuitive, Versus a Systematic, Mode of Thought Produces Greater Empathic Accuracy', *Journal of Personality and Social Psychology* 111, no. 5 (2016): 674–85, accessed on 4 January 2023 and viewable at: <https://www.apa.org/pubs/journals/releases/psp-pspi0000063.pdf>.

- 4 Evan Mimms, '86% of patients say wearable devices improve health outcomes, according to software advice research', Businesswire (2 March 2022), accessed on 11 October 2023, viewable at:
<https://www.businesswire.com/news/home/20220302005174/en/86-of-Patients-Say-Wearable-Devices-Improve-Health-Outcomes-According-to-Software-Advice-Research>.
- 5 'Wearable technology market', Markets and Markets (April 2021), accessed on 19 July 2023, viewable at:
<https://www.marketsandmarkets.com/Market-Reports/wearable-electronics-market-983.html>.
- 6 Michael Luca, 'Reviews, Reputation, and Revenue: The Case of Yelp.com', Harvard Business School (2016), accessed on 11 October 2023, viewable at:
<https://www.hbs.edu/faculty/Pages/item.aspx?num=41233>.
- 7 'Global Issues: Population', United Nations, accessed on 19 July 2023, viewable at: <https://www.un.org/en/global-issues/population>.
- 8 'Current World Population', Worldometer (16 July 2023), accessed on 19 July 2023, viewable at:
<https://www.worldometers.info/world-population/>.
- 9 Larry Dignan, 'IBM launches Watson tools for agriculture', ZDNet (21 May 2019), accessed on 14 October 2023, viewable at:
<https://www.zdnet.com/article/ibm-launches-watson-tools-for-agriculture/>; 'IBM AI and Cloud Technology helps agriculture industry improve the world's food and crop supply', IBM (22 May 2019), accessed on 14 October 2023, viewable at:
<https://newsroom.ibm.com/2019-05-22-IBM-AI-and-Cloud-Technology-Helps-Agriculture-Industry-Improve-the-Worlds-Food-and-Crop-Supply>.
- 10 'Big Data and Agriculture: A Complete Guide', Talend, accessed on 19 July 2023, viewable at:
<https://www.talend.com/resources/big-data-agriculture/>.

Chapter 23: Information and Wisdom

- 1 'Sumerian Religion', RootsHunt, accessed on 11 October 2023, viewable at:
<https://rootshunt.com/aryans/indusvalleycivilization/sumerianmyth/sumerianmyth.htm>.
- 2 'What does silicon valley's bank's collapse mean for the financial system?', Livemint (11 March 2023), accessed on 11 October 2023, viewable at: <https://www.livemint.com/market/what-does-silicon-valley-bank-s-collapse-mean-for-the-financial-system-11678512462626.html>.
- 3 R.L. Ackoff, 'From data to wisdom presidential address to ISGSR', *Journal of Applied Systems Analysis* 16, (July 1989): 3–9.
- 4 T.S. Elliot, 'The Rock', Wisdom Portal, accessed on 11 October 2023, viewable at:
<https://www.wisdomportal.com/Technology/TSEliot-TheRock.html>.
- 5 'Think fast, think slow', Wikipedia, accessed on 11 October 2023 and available at:
https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow.
- 6 Linda Adams, 'Learning a new skill is easier said than done', Gordon Training International, accessed on 11 October 2023, viewable at: <http://www.gordontraining.com/free-workplace-articles/learning-a-new-skill-is-easier-said-than-done/>.
- 7 Monika Ardelt and Bhavna Sharma, 'Linking wise organizations to wise leadership, job satisfaction, and well-being', *Frontiers* (19 November 2021), accessed on 11 October 2023, viewable at:
<https://www.frontiersin.org/articles/10.3389/fcomm.2021.685850>; Gary Yukl and Rubina Mahsud, 'Why flexible and adaptive leadership is essential', Consulting Psychology Journal Practice and Research (June 2010), accessed on 11 October 2023, viewable at: https://www.researchgate.net/publication/232567495_Why_flexible_and_adaptive_leadership_is_essential; Bernard McKenna and David Rooney, 'Wise Leadership', *The Cambridge Handbook of Wisdom* (Cambridge University Press, 15 March 2019), accessed on 11 October 2023, viewable at:
<https://www.cambridge.org/core/books/abs/cambridge-handbook-of-wisdom/WISE-LEADERSHIP>

[of-wisdom/wise-leadership/BF97FC7AE757F80F371416B55D1DC642.](#)

Chapter 24: Data Sharing vs Data Privacy

- 1 Drew Donnelly, 'China Social Credit system explained – what is it & how does it work?', Horizons (28 September 2023), accessed on 12 October 2023, viewable at: <https://joinhorizons.com/china-social-credit-system-explained/>; Nicole Kobie, 'The complicated truth about China's social credit system', Wired UK (6 July 2019), accessed on 12 October 2023, viewable at: <https://wired.co.uk/article/china-social-credit-system-explained>; Eunsun Cho, 'The Social Credit system: not just another Chinese idiosyncrasy', Princeton University (1 May 2020), accessed on 12 October 2023, viewable at: <https://jpiia.princeton.edu/news/social-credit-system-not-just-another-chinese-idiosyncrasy>.
- 2 'Embracing Innovation in Government: Global Trends 2018, Case Study AADHAR', OECD, accessed on 12 October 2023, viewable at: <https://www.oecd.org/gov/innovative-government/India-case-study-UAE-report-2018.pdf>.
- 3 ABC News, 'Senator catches Mark Zuckerberg off guard by asking which hotel he stayed in', YouTube (11 April 2018), accessed on 12 October 2023, viewable at: <https://www.youtube.com/watch?v=TX8MSZy5I3I>.
- 4 Mathieu Rosemain, 'France fines Google \$57 million for European privacy rule breach', Reuters (22 January 2019), accessed on 12 October 2023, viewable at: <https://www.reuters.com/article/us-google-privacy-france-idUSKCN1PF208>.
- 5 Ilker Koksal, 'Twitter admits to exploiting users' Personal Data', *Forbes* (1 November 2019), accessed on 12 October 2023, viewable at: <https://www.forbes.com/sites/ilkerkoksal/2019/11/01/twitter-admits-to-exploiting-users-personal-data/>.
- 6 Nate Raymond and Joseph Ax, 'Ex-Morgan Stanley adviser pleads guilty in connection with data breach', Reuters (22 September 2015), accessed on 19 July 2023, viewable at:

<https://www.reuters.com/article/us-morgan-stanley-breach-plea-idUSKCN0RL22920150921>.

- 7 John S. Hollywood, 'CPD's 'Heat List' and the Dilemma of Predictive Policing', RAND (21 September 2016), accessed on 12 October 2023, viewable at: <https://www.rand.org/blog/2016/09/cpds-heat-list-and-the-dilemma-of-predictive-policing.html>.
- 8 Lauren Feiner and Christina Wilkie, 'White House endorses new Senate TikTok bill, urges Congress to pass it 'quickly'', CNBC (7 March 2023), accessed on 12 October 2023, viewable at: <https://www.cnbc.com/2023/03/07/white-house-endorses-senate-tiktok-bill-urges-congress-to-pass-soon.html>.
- 9 'Minority Report', IMDB (2002), accessed on 12 October 2023, viewable at: <https://www.imdb.com/title/tt0181689/>.
- 10 'Convention 108 and Protocols', Council of Europe, accessed on 12 October 2023, viewable at: <https://www.coe.int/en/web/data-protection/convention108-and-protocol>.
- 11 Legal Writing Team, 'What's Data privacy law in your country?', Privacy Policies (14 March 2023), accessed on 12 October 2023, viewable at: <https://www.privacypolicies.com/blog/privacy-law-by-country/>.
- 12 'Global Privacy Assembly', accessed on 12 October 2023, viewable at: <https://globalprivacyassembly.org/>.
- 13 Carole Cadwalladr and Emma Graham-Harrison, 'Revealed: 50 million Facebook profiles harvested for Cambridge Analytica', *The Guardian* (17 March 2018), accessed on 12 October 2023, viewable at: <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election> major data breach'.
- 14 Brooke Auxier et al., 'Americans and privacy: concerned, confused and feeling lack of control over their personal information', Pew Research (15 November 2019), accessed on 12 October 2023, viewable at: <https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/>.

15 Akindare Okunola, 'Your Data Is Our Data: A Truecaller Breakdown', Tech Cabal (2 May 2018), accessed on 12 October 2023, viewable at: <https://techcabal.com/2018/05/02/your-data-is-our-data-a-truecaller-breakdown/>.

16 Emmanuel Moses Temidayo, 'Why you should consider uninstalling your Truecaller App', Dignited (1 August 2020), accessed on 20 July 2023, viewable at: <https://www.dignited.com/55465/why-you-should-consider-uninstalling-your-truecaller-app/>.

Chapter 25: Digital Engagement vs Mental Health

1 'Over a third of Brits feel stressed every day due to data overload', ESRI UK (29 September 2022), accessed on 13 October 2023, viewable at: <https://www.esriuk.com/en-gb/news/press-releases/uk/24-stress-map>.

2 Simon Kemp, 'Kids as young as 8 are using social media more than ever, study finds', *New York Times* (24 March 2022), accessed on 17 November 2023, viewable at: <https://www.nytimes.com/2022/03/24/well/family/child-social-media-use.html>.

3 Ibid.

4 Stacy Jo Dixon, 'Social network penetration worldwide from 2018 to 2027', Statista (July 2023), accessed on 20 July 2023, viewable at: <https://www.statista.com/statistics/260811/social-network-penetration-worldwide/>.

5 'OTT Video - Worldwide', Statista (July 2023), accessed on 20 July 2023, downloadable from: <https://www.statista.com/outlook/amo/media/tv-video/ott-video/worldwide>.

6 Daniela Coppola, 'E-commerce as percentage of total retail sales worldwide from 2015 to 2026', Statista (29 August 2023), accessed on 4 January 2024, viewable at: <https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/>.

- 7 Claire McCarthy, 'Can cell phone use cause ADHD?', Harvard Health Publishing (31 July 2018), accessed on 20 July 2023, viewable at: <https://www.health.harvard.edu/blog/can-cell-phone-use-cause-adhd-2018073114375>.
- 8 Aaron Smith, 'U.S. smartphone use in 2015', Pew Research (1 April 2015), accessed on 20 July 2023, viewable at: <https://www.pewresearch.org/internet/2015/04/01/us-smartphone-use-in-2015/>.
- 9 'Spiritual Balance – a necessity in today's material world', Nitin's Fundas (26 August 2013), accessed on 20 July 2023, viewable at: <https://nseth71.blogspot.com/2013/08/spiritual-balance-necessity-in-todays.html>.

Chapter 26: Data Collaboration for a Better World

- 1 'Where the Web was born', CERN, accessed on 13 October 2023, viewable at: <https://home.cern/science/computing/birth-web/short-history-web>.
- 2 'The Sustainable Development Agenda', United Nations, accessed on 13 October 2023, viewable at: <https://www.un.org/sustainabledevelopment/development-agenda/>.
- 3 Adam Mahdi et al., 'OxCOVID19 Database, a multimodal data repository for better understanding the global impact of COVID-19', *Nature* (29 April 2021), accessed on 13 October 2023, viewable at: <https://www.nature.com/articles/s41598-021-88481-4>.
- 4 Ivana Kottasová and Krystina Shveda, 'Health Genomic sequencing is crucial in the battle against the coronavirus. These countries do it well', CNN (9 December 2021), accessed on 13 October 2023, viewable at: <https://edition.cnn.com/2021/12/09/health/coronavirus-genomic-sequencing-intl-cmd/index.html>.
- 5 'WHO names new covid variant Omicron, cautions against travel measures', VOA News (26 November 2021), accessed on 13 October 2023, viewable at: <https://www.voanews.com/a/who->

[designates-covid-variant-found-in-south-africa-as-of-concern/6329408.html](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8600000/).

- 6 Jocelyn Solis-Moreira, 'How did we develop a COVID-19 vaccine so quickly?', Medical News Today (13 November 2021), accessed on 13 October 2023, viewable at:
<https://www.medicalnewstoday.com/articles/how-did-we-develop-a-covid-19-vaccine-so-quickly>.
- 7 Charles Schmidt, 'Genetic engineering could make a covid-19 vaccine in months rather than years', Scientific American (1 June 2020), accessed on 21 July 2023, viewable at:
<https://www.scientificamerican.com/article/genetic-engineering-could-make-a-covid-19-vaccine-in-months-rather-than-years1/>.
- 8 Clare Watson, 'Rise of the preprint: How rapid data sharing during COVID-19 has changed science forever', *Nature* (14 January 2022), accessed on 13 October 2023, downloadable from:
<https://www.up.ac.za/media/shared/538/Research%20services%20and%20support/Documents%20and%20References/watson-2022.zp214736.pdf>.
- 9 'Clinical research metadata repository', ECRIN, accessed on 13 October 2023, viewable at: <https://ecrin.org/clinical-research-metadata-repository>.
- 10 Megan Cerullo, 'Smartwatches can help detect COVID-19 days before symptoms appear', CBS News (15 January 2021), accessed on 13 October 2023, viewable at:
<https://www.cbsnews.com/news/covid-symptoms-smart-watch/>.
- 11 'A Brief History of MOOCs', McGill Association of University Teachers, accessed on 13 October 2023, viewable at:
<https://www.mcgill.ca/maut/news-current-affairs/moocs/history>.
- 12 Ibid.
- 13 Andrea Mercado, 'Coursera Statistics: The Growth of e-Learning in 2023', Skillademia (3 April 2023), accessed on 15 October 2023, viewable at: <https://www.skillademia.com/statistics/coursera-statistics/#:~:text=Coursera%20has%20a%20community%20of,students%20on%20Coursera%20are%20male>.
- 14 'Association for Molecular Pathology et al. v. Myriad Genetics, Inc. et al.', UNCTAD (13 June 2013), accessed on 13 October 2023,

viewable at:

<https://unctad.org/ippcaselaw/sites/default/files/ippcaselaw/2020-12/Association%20for%20Molecular%20Pathology%20et%20al.%20v.%20Myriad%20Genetics%2C%20U.S.%20Supreme%20Court%202013.pdf>.

- 15 Ben Wodecki, 'IDC: China set to more than double AI spending by 2026', AI Business (12 October 2022), accessed on 13 October 2023, viewable at: <https://aibusness.com/verticals/idc-china-set-to-more-than-double-ai-spending-by-2026>.
- 16 'Japan pushing ahead with Society 5.0 to overcome chronic social challenges', UNESCO (21 February 2019), accessed on 21 July 2023, viewable at: <https://www.unesco.org/en/articles/japan-pushing-ahead-society-50-overcome-chronic-social-challenges>.
- 17 Yonhap, 'S. Korea to invest over 20 Tr won in data, network, AI sectors', Korea Herald (25 March 2022), accessed on 21 July 2023, viewable at: <https://www.koreaherald.com/view.php?ud=20220325000326>.
- 18 'Economic Survey 2023: 135.2 crore Aadhaar numbers generated till November 2022', Moneycontrol (31 January 2023), accessed on 21 July 2023, viewable at:
<https://www.moneycontrol.com/news/business/budget/economic-survey-2023-135-2-crore-aadhaar-numbers-generated-till-november-2022-9971821.html>.
- 19 'Frequently asked questions', Global Partnership on Artificial Intelligence (November 2021), accessed on 13 October 2023, downloadable from: <https://www.gpai.ai/about/gpai-faq.pdf>.

Chapter 27: Data as a Source of National Competitive Advantage

- 1 'Global education monitoring report 2023', UNESCO (2023), accessed on 12 October 2023, downloadable from:
<https://unesdoc.unesco.org/ark:/48223/pf0000385723>.
- 2 Gregory C. Allen (Congressional testimony), 'China's Pursuit of Defense Technologies: Implications for U.S. and Multilateral Export Control and Investment Screening Regimes', Center for Strategic

and International Studies (13 April 2023), accessed on 12 October 2023 and viewable at: <https://www.csis.org/analysis/chinas-pursuit-defense-technologies-implications-us-and-multilateral-export-control-and>.

- 3 Stephen Witt, 'The Turkish drone that changed the nature of warfare', *New Yorker* (9 May 2022), accessed on 12 October 2023, viewable at:
<https://www.newyorker.com/magazine/2022/05/16/the-turkish-drone-that-changed-the-nature-of-warfare2>.
- 4 Joyce Chen, 'BTS: Successful social media strategy of the legendary K-POP group', Medium (14 February 2020), accessed on 12 October 2023, viewable at: <https://medium.com/digital-society/bts-successful-social-media-strategy-of-the-legendary-k-pop-group-5d29b7eb09dd>.
- 5 'e-Estonia, the information society since 1997', Center for Public Impact (2 September 2019), accessed on 12 October 2023, viewable at: <https://www.centreforpublicimpact.org/case-study/e-estonia-information-society-since-1997>.
- 6 Ibid.
- 7 Allegra Crahay, 'eID and e-Signature in cross-border situations, the Estonian experience', European Commission (15 April 2016), accessed on 12 October 2023, viewable at:
<https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/document/eid-and-e-signature-cross-border-situations-estonian-experience-estonian-eid-and-e-signature>.
- 8 'Estonia PM: Country saves 2% of GDP by going digital', International Peace Institute (3 May 2016), accessed on 12 October 2023, viewable at:
<https://www.ipinst.org/2016/05/information-technology-and-governance-estonia#4>.
- 9 'Estonian blockchain technology FAQs', e-Estonia (March 2020), accessed on 12 October 2023, viewable at: <https://e-estonia.com/wp-content/uploads/2020mar-nochanges-faq-a4-v03-blockchain-1-1.pdf>.

- 10 Anita Tuula, Kristiina Sepp and Daisy Volmer, 'E-solutions in Estonian community pharmacies: A literature review', National Library of Medicine (18 July 2022), accessed on 12 October 2023, viewable at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9301098/>.
- 11 'e-Governance', e-Estonia, accessed on 12 October 2023, downloadable from: <https://e-estonia.com/solutions/e-governance/data-embassy/>.
- 12 Siim Sikkut, 'AI - "kratt" strategy', e-Estonia, accessed on 21 December 2023, viewable at: <https://e-estonia.com/wp-content/uploads/2020-april-facts-ai-strategy.pdf>.
- 13 Harry Lynch, 'How Middle Eastern geopolitics is boosting Israel's cybersecurity scene', Disruption Banking (20 September 2022), accessed on 12 October 2023, viewable at:
<https://www.disruptionbanking.com/2022/09/20/how-middle-eastern-geopolitics-is-boosting-israels-cybersecurity-scene/>.
- 14 'CyberArk software market cap', Stock Analysis, accessed on 12 October 2023, viewable at:
<https://stockanalysis.com/stocks/cybr/market-cap/>; 'Check Point Software Technologies Ltd market cap', YCharts, accessed on 12 October 2023, viewable at:
https://ycharts.com/companies/CHKP/market_cap.
- 15 Richard Behar, 'Inside Israel's secret startup machine', *Forbes* (11 May 2016), accessed on 12 October 2023, viewable at:
<https://www.forbes.com/sites/richardbehar/2016/05/11/inside-israels-secret-startup-machine/>.
- 16 Tali Tsipori, '8200 graduates aren't like 23 year-olds in Texas or Norway', Globes (5 June 2017), accessed on 12 October 2023, viewable at: <https://en.globes.co.il/en/article-8200-graduates-are-not-like-23-year-olds-in-texas-or-norway-1001191294>.
- 17 'Promoting digital innovation to deliver value to Korean citizens - OECD Digital Government Index 2019', OECD, accessed on 12 October 2023, viewable at:
<https://www.oecd.org/country/korea/digital-government>.
- 18 'Utilizing Big Data to solve urban issues: The case of Seoul', Seoul Metropolitan Government, accessed on 12 October 2023,

downloadable from:

https://www.thegpsc.org/sites/gpsc/files/partnerdocs/seoul_utilizing_big_data_to_solve_urban_issues - the case of seoul.pdf.

19 Ibid.

20 'Seoul City builds the nation's first urban problem-solving simulation Digital Twin S-Map', Smart City Korea (1 April 2021), accessed on 12 October 2023, viewable at:

<https://smartcity.go.kr/en/2021/04/01/%EC%84%9C%EC%9A%B8%EC%8B%9C-%EC%8F%84%EC%8B%9C%EB%AC%B8%EC%A0%9C%ED%95%B4%EA%B2%B0-%EC%8B%9C%EB%AE%AC%EB%A0%88%EC%9D%B4%EC%85%98-%EB%94%94%EC%A7%80%ED%84%B8-%ED%8A%B8%EC%9C%88-s-map-%EC%A0%84/>.

21 PTI, 'UIDAI rolls out new security mechanism for robust fingerprint-based Aadhaar authentication', *Times of India* (28 February 2023), accessed on 12 October 2023, viewable at:
<https://timesofindia.indiatimes.com/india/uidai-rolls-out-new-security-mechanism-for-robust-fingerprint-based-aadhaar-authentication/articleshow/98288692.cms>.

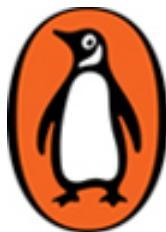
22 'Economic Survey 2023: UPI accounted for 52% of India's total digital transactions in FY22', Moneycontrol (31 January 2023), accessed on 12 October 2023, viewable at:
<https://www.moneycontrol.com/news/business/economic-survey-2023-upi-accounted-for-52-of-indias-total-digital-transactions-in-fy22-9970741.html>.

23 Jessica Rajan, 'UPI crosses 10 billion transactions in September for second straight month', *Economic Times* (2 October 2023), accessed on 12 October 2023, viewable at:
<https://economictimes.indiatimes.com/tech/technology/upi-crosses-10-billion-transactions-in-september-for-second-straight-month/articleshow/104101957.cms>.

24 Laxitha Mundhra, 'India to provide UPI to the world; in talks with 30 countries: IT Minister', Inc42 (5 July 2022), accessed on 12 October 2023, viewable at: <https://inc42.com/buzz/india-to-provide-upi-to-the-world-in-talks-with-30-countries-it-minister/>.



Scan QR code to access the Penguin Random House India website



THE BEGINNING

Let the conversation begin...

Follow the Penguin [Twitter.com@penguinbooks](https://twitter.com/penguinbooks)

Keep up-to-date with all our stories [YouTube.com/penguinbooks](https://www.youtube.com/penguinbooks)

Pin 'Penguin Books' to your [Pinterest](#)

Like 'Penguin Books' on [Facebook.com/penguinbooks](https://facebook.com/penguinbooks)

Find out more about the author and
discover more stories like this at [Penguin.co.in](https://penguin.co.in)

PENGUIN BUSINESS

USA | Canada | UK | Ireland | Australia

New Zealand | India | South Africa | China | Singapore

Penguin Business is part of the Penguin Random House group of companies
whose addresses can be found at global.penguinrandomhouse.com



Penguin
Random House
India

This collection published 2024

Copyright © Nitin Seth 2024

The moral right of the author has been asserted

Jacket images © Sparsh Raj Singh

This digital edition published in 2024.

e-ISBN: 978-9-357-08784-1

This book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.