McKinsey
Digital

# In search of cloud value

Can generative AI transform cloud ROI?

November 2023

# In search of cloud value

Can generative AI transform cloud ROI?

# Contents

# Overview

Established companies face a quandary as they look to exploit cloud: as attractive as the benefits may be, the scale of change and investments required to adopt cloud platforms make generating an attractive return on investment (ROI) a challenge. But generative AI may significantly shift that value equation. It has the potential to dramatically reduce the investment and time needed to adopt cloud *and* generate new value by unlocking new business and tech use cases.

Cloud platforms have the potential to enable massive value in every sector. Enterprise technology users in the Global 2000 could increase annual EBITDA by more than $3 trillion by 2030 by increasing IT productivity, creating new value, and opening up new businesses and business models. While the possible impact varies by sector, adopting cloud represents an opportunity for the average company to increase profitability by 20 to 30 percent.

Many digital-native companies are already taking full advantage of this opportunity. Nearly one-third of the EBITDA value gain over the past decade in the S&P 500 has come from just eight digital-native companies that utilized cloud-like infrastructure. While many incumbent enterprises aspire to do the same, results have generally been short of expectations. Only 10 percent of companies report they are capturing value at scale from cloud investments.

These returns have led to low adoption rates, especially at the core of the enterprise. As a median, large companies run only 15 to 20 percent of their applications in cloud, even when they have been running cloud programs for years and even after they account for the use of software-as-a-service (SaaS) products. Nor is it clear that there has been a dramatic uptick in adoption over the past year—companies with cloud programs profiled by McKinsey have increased their cloud adoption by only 5 to 10 percent over the past 12 months. Aspirations for cloud adoption, nevertheless, remain high. Almost all of the more than 80 enterprises McKinsey profiled for its CloudSights database (see sidebar “About CloudSights: ROI engine” in Part I) aspire to run the majority of their applications in public cloud within five to seven years; more than two-thirds aspire to run 80 percent of their systems in cloud.

What is the reason for this disconnect between aspiration and reality? Getting value from public cloud, it turns out, is complicated. Companies have spent the past several decades building enterprise technology organizations, processes, and architectures designed to work for on-premises environments. Much of that needs to change.

An effective cloud strategy requires not just changing servers but also refining and sometimes reinventing how technology is developed, operated, and managed. Cloud requires different application architectures, new types of infrastructure services, different operational capabilities—all of which require time and investment. Merely remediating business applications so they run in a secure, scalable, efficient, and resilient way in cloud can cost hundreds of millions of dollars for medium-size enterprise technology organizations and billions of dollars for larger ones—pushing breakeven points for even well-run cloud programs to five to seven years. Besides managing the technical complexities, companies need to essentially implement a set of operational and organizational changes as well to get the benefits of technology, as highlighted in McKinsey’s recent book *Rewired*.[1]

In response to these realities, many companies have moved slowly, spreading their investments over many years to tap existing technology, refresh funding, and reduce “bubble costs” in their IT budgets.

[1] Eric Lamarre, Kate Smaje, and Rodney Zemmel, *Rewired: The McKinsey Guide to Outcompeting in the Age of Digital and AI*, New York: Wiley, 2023.

Generative AI could transform the cloud investment-and-return equation. When McKinsey gathered nearly 80 CTOs and cloud program leaders together this fall, we heard that many believe generative AI may be a disruptor that transforms ROI dynamics for cloud programs and accelerates cloud adoption.

There are two elements to this opportunity. One is using cloud to support generative AI initiatives. With its massive calls on compute, storage, and networking, generative AI needs cloud to scale. Generative AI's complexity, moreover, requires implementation via scalable enterprise cloud platforms rather than via disconnected pilots and initiatives run by individual development teams.

The second element of opportunity is using generative AI capabilities to accelerate cloud programs. Currently, remediating some applications to run effectively in cloud typically requires investments equal to several years' worth of support and maintenance costs. Early efforts to apply generative AI to application remediation and migration have indicated a 40 percent reduction in time and investment required, though much work still is needed to understand how the improvements apply for different types of applications. The creation of end-to-end, generative AI–enabled workflows will create incentives for companies to migrate existing transactional applications to cloud.

The findings in this report paint a detailed picture of where cloud's overall value lies and what it takes for companies to capture their fair share of it. Following are a few of the highlights:

— Cloud can generate about $3 trillion in EBITDA by 2030.

— The value cloud generates from enabling businesses to innovate is worth more than *five times* what is possible by simply reducing IT costs.

— Across sectors, the potential EBITDA uplift from cloud by 2030 averages 20 to 30 percent over the projected baseline, but it varies significantly by sector, with the greatest potential in high tech and the least in electric utilities.

— Asian companies have the most to gain from cloud, with $1.2 trillion in EBITDA by 2030 at stake, driven by a higher rate of baseline revenue growth and more room to grow. American institutions stand to capture about $1.1 trillion in cloud value, while European institutions may have a somewhat smaller opportunity of $773 billion, primarily due to regulatory headwinds.

— An average company adopting cloud today could achieve 180 percent ROI in business benefit, although few are getting close to these returns.

— Only 10 percent of companies have fully captured cloud's potential value, while another 50 percent are starting to capture it, and the remaining 40 percent have seen no material value.

— Nearly 40 percent of companies now say "business value" determines which applications move to cloud—up from 27 percent in 2021 and 2022.

— Companies that have captured the most ROI consistently do three things well: work closely with business leaders to focus on high-value business cases, build a robust cloud foundation, and adopt a product-oriented operating model.

- Lost value in cloud programs comes from three primary sources: unrealized use cases (focusing on IT savings rather than new value), cloud sprawl (redundant cloud foundations), and stalled adoption (breakeven generally comes at around 50 percent cloud adoption). Taken together, these three factors can completely erase the benefits cloud can provide—and even destroy value.

- Generative AI can add 75 to 110 percentage points of incremental ROI to cloud programs through three key benefits:  unlocking new business use cases; reducing the time and cost of application remediation and migration (early results suggest 40 percent lower time and costs); and increasing the productivity of application development and infrastructure teams on cloud.

- Going forward, companies can build generative AI into their cloud programs in a number of ways: incorporating generative AI–enabled business use cases; accelerating migration of on-premises transactional systems to build end-to-end generative AI–enabled customer journeys; and using generative AI to transform the ROI of application remediation and migration. To ensure the safe and efficient operation of generative-AI capabilities in cloud, they should be built into the entire cloud program, from foundational platforms to FinOps tooling and security capabilities.

This report is organized around the top three questions business leaders need to answer to get more value from their cloud programs:

**Part I: What is the true value of cloud?** This section breaks down all the sources of value from cloud across three areas: IT productivity, business innovation, and advanced technologies, including where generative AI can expand the value from cloud adoption. We look at how this value plays out across industries and geographies. Part I also digs into the root causes of why many cloud programs struggle.

**Part II: How can my company maximize its cloud ROI?** This part of the report explores how organizations can balance cloud investments against the expected benefits they will generate (including the transformation of ROI dynamics through generative AI), enabling them to plan for and prioritize their migration journeys. Part II also lays out some of the different pathways organizations may choose for their cloud programs and how each path might affect the ROI trajectory.

**Part III: What actions should we take?** This section turns to practicalities and lays out the ten essentials—actions that institutions can take right now to build healthy cloud programs and maximize ROI. We organize these ten actions around three broad areas:

- **Discover** the full value.
- **Solve** critical technical problems to enable the business to capture the value.
- **Deliver** the necessary organizational change. We also show what cloud programs need to do to support generative AI capabilities.

## Part I

# What is the true value of cloud?

The $3 trillion opportunity

**Our latest analysis estimates** that cloud could generate about $3 trillion in EBIDTA by 2030 (see sidebar "Methodology for sizing the value of cloud through 2030"). The sources of that value fall into three dimensions (Exhibit 1):

1. **Rejuvenate** refers to creating value by optimizing IT costs, including infrastructure, application maintenance, and development. It also includes reduced costs of business downtime due to improved IT resiliency. Examples of IT cost reduction include exiting a costly data center, improving server utilization through auto-scaling cloud infrastructure, automating application maintenance tasks, and increasing developer productivity through reuse of cloud capabilities. Improved IT resiliency can come from reduced incidents and faster recovery due to more-standard, redundant, and self-healing cloud infrastructure. Rejuvenate could generate approximately $570 billion in EBITDA uplift for companies by 2030.[2]

2. **Innovate** refers to creating value by using cloud to digitize core business operations, drive new business innovation, and generate new growth across 700 use cases our research identified. Examples include value generated from new and enhanced use cases in analytics, IoT, and automation; accelerated product development; and harnessing cloud's scalability in compute and storage capacity. A pharma company looking to get value by innovating, for example, might use artificial intelligence to streamline clinical trials or help accelerate the discovery of new drugs. Innovate accounts for about $2.5 trillion in our estimate of total cloud value by 2030.[3]

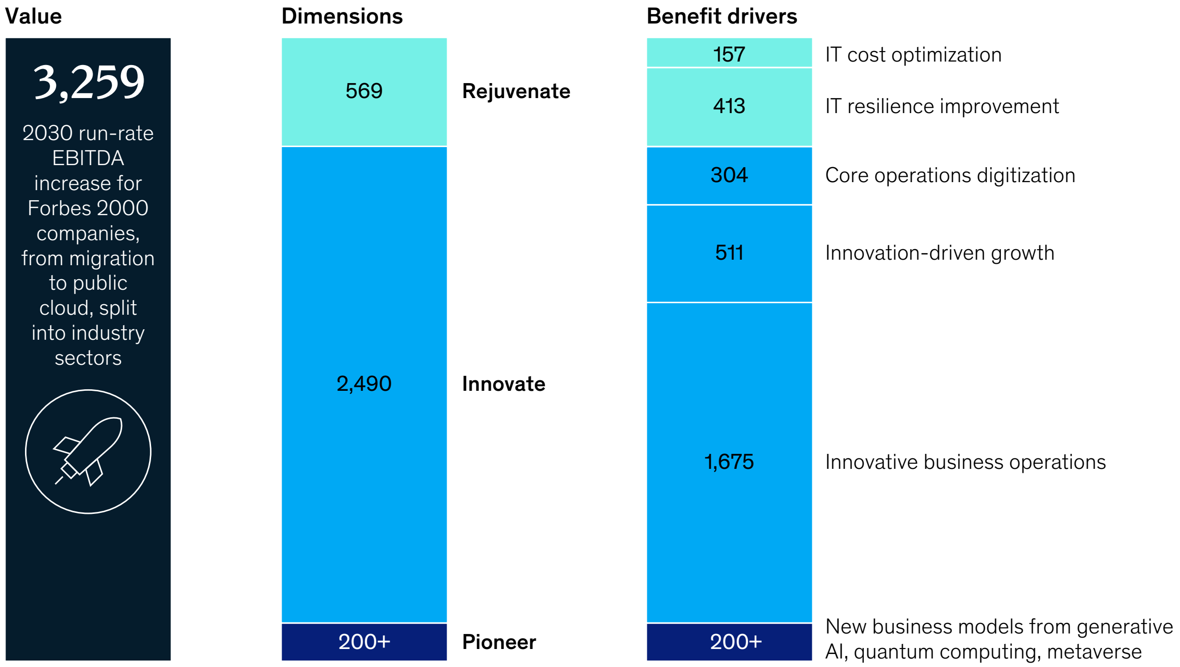
3. **Pioneer** refers to creating value through early and rapid experimentation and scaling new technologies, such as generative AI, quantum, and immersive reality. We expect a smaller set of companies to gain outsized value in Pioneer, based on the fact that, from 2013–23, the EBITDA gains of only eight "born-digital" organizations accounted for $260 billion of the S&P 500's total EBITDA growth of $960 billion. These "born-digital" companies pioneered the biggest technology trends in the past decade and leveraged cloudlike technology infrastructure to rapidly innovate and achieve scale, disrupting traditional business models. The great news is that this type of highly scalable infrastructure is now available for all organizations through public cloud.

[2] McKinsey D2020 proprietary IT cost benchmarking; 2019 third-party cloud economics benchmarking; MGI research; expert interviews; team analysis.
[3] Ibid.

Exhibit 1

## The sources of estimated cloud value fall into three dimensions: Rejuvenate, Innovate, and Pioneer.

**Estimated additional run-rate EBITDA of Forbes 2000 companies in 2030,** $ billion



Note: Figures may not sum to totals, due to rounding.
Source: Third-party cloud economics benchmarking, 2019; McKinsey D2020 proprietary IT cost benchmarking; McKinsey Global Institute research; expert interviews; team analysis

## The opportunities are massive, but vary significantly by sector

Some industries will have an opportunity to generate more value from cloud than others (Exhibit 2). High tech, oil and gas, retail, healthcare systems and services, insurance, and banking are positioned to generate the most value by 2030, as measured by EBITDA impact, although almost all industries across the Forbes 2000 show the potential to increase EBITDA by an average of more than 20 percent.[4]

In our model, five major factors drive the differences in potential EBITDA impact from cloud by industry:

- **The industry's revenue share in the Forbes 2000** (the total revenues of companies in a given industry listed among the top 2,000 companies globally by revenue)
- **IT spend intensity** (the percentage of revenues a given industry spends on IT)
- **AI potential** (the potential EBITDA impact of cloud-enabled AI use cases)
- **IoT potential** (the EBITDA impact of cloud-enabled IoT use cases)
- **Automation potential** (the EBITDA impact of cloud-enabled automation use cases)

[4] "Projecting the global value of cloud: $3 trillion is up for grabs for companies that go beyond adoption," McKinsey, November 28, 2022.

## Methodology for sizing the value of cloud through 2030

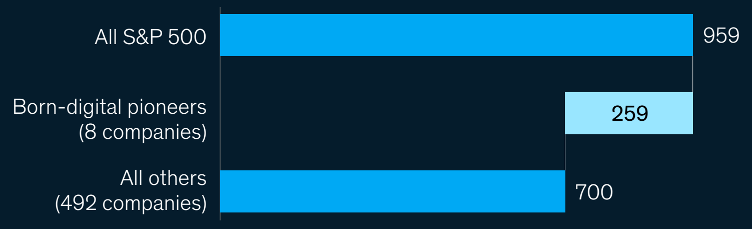
To quantify the total potential value for the Rejuvenate and Innovate dimensions, we conducted detailed analyses based on three reports from the McKinsey Global Institute (MGI); McKinsey D2020 benchmarking for IT spending structure based on more than 1,000 IT diagnostics worldwide; and independent third-party surveys of more than 1,000 organizations that have adopted cloud to pursue potential gains in operational efficiency. In applying the MGI research, we assessed more than 700 use cases across 20 subindustries. We also utilized IHS Markit industry growth rates to establish baselines for 2030 financial performance of the Forbes Global 2000 without cloud-based EBITDA lifts.[1]

To quantify the potential value from the Pioneer category, we looked at the EBITDA performance of the "born-digital" companies that pioneered the key tech trends starting ten years ago. We identified eight exemplary companies that generated almost $260 billion of EBITDA growth, representing nearly a third of all EBITDA growth for the S&P 500 (exhibit).

Exhibit

### Eight 'digital pioneers' account for 27 percent of the past decade's EBITDA growth in the S&P 500.

**Net EBITDA growth of S&P 500 companies from 2013–23,** $ billion



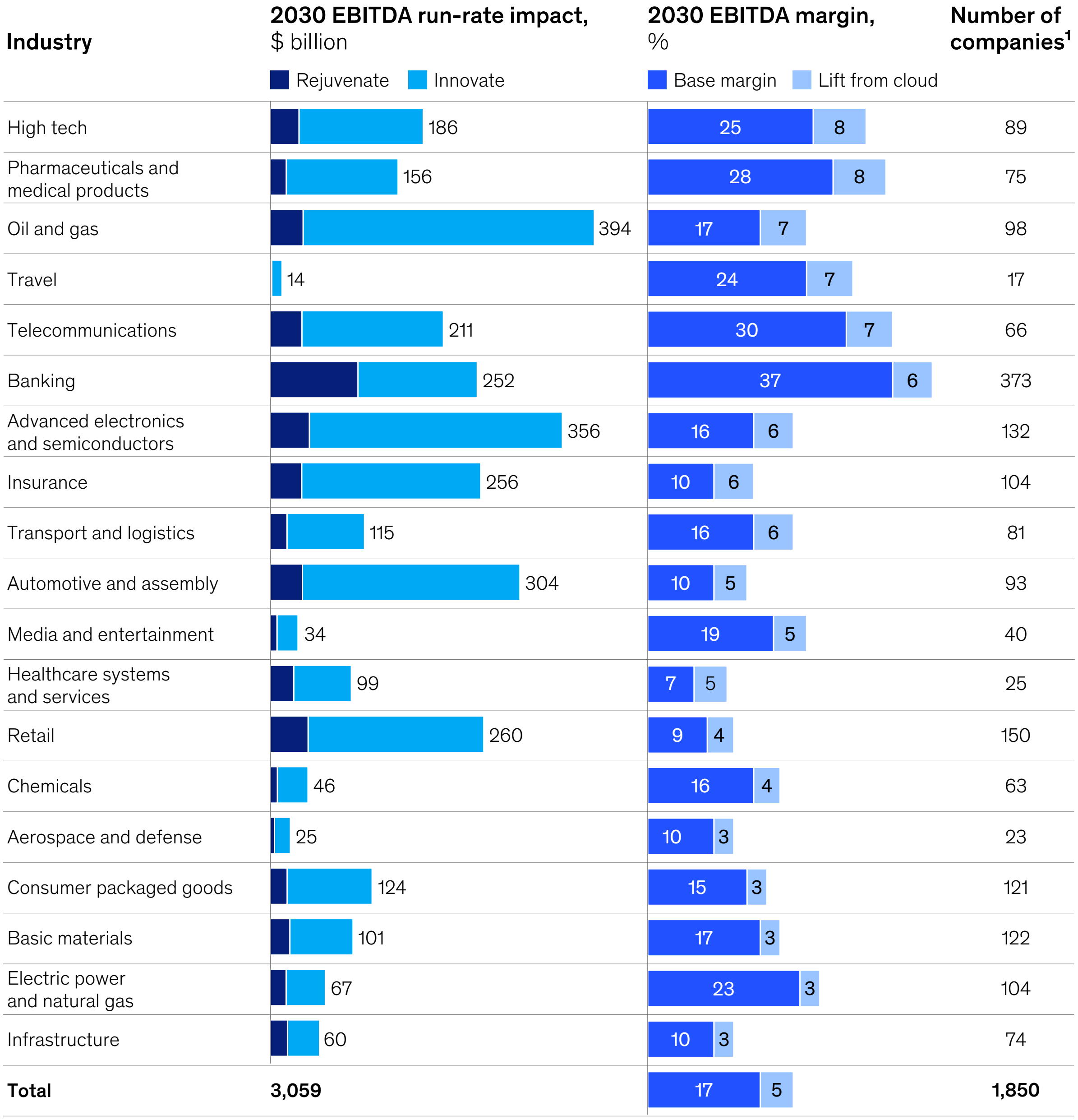
**8 born-digital companies** generated **27% of the total net EBITDA growth** across the S&P 500 by pioneering the biggest tech trends in the past 10 years

We have chosen to use EBITDA, as opposed to alternative metrics such as EBIT or operating income, because it is the most common cross-sector metric for business profitability that is not impacted by the timing and magnitude of capital investments, interest, and taxes.

[1] See the methodology in "Projecting the global value of cloud: $3 trillion is up for grabs for companies that go beyond adoption," McKinsey, November 28, 2022.

Exhibit 2

## Capture of cloud's economic value is expected to differ by industry.



| Industry | 2030 EBITDA run-rate impact, $ billion | 2030 EBITDA margin, %: Base margin | 2030 EBITDA margin, %: Lift from cloud | Number of companies[1] |
|---|---|---|---|---|
| | Rejuvenate / Innovate | | | |
| High tech | 186 | 25 | 8 | 89 |
| Pharmaceuticals and medical products | 156 | 28 | 8 | 75 |
| Oil and gas | 394 | 17 | 7 | 98 |
| Travel | 14 | 24 | 7 | 17 |
| Telecommunications | 211 | 30 | 7 | 66 |
| Banking | 252 | 37 | 6 | 373 |
| Advanced electronics and semiconductors | 356 | 16 | 6 | 132 |
| Insurance | 256 | 10 | 6 | 104 |
| Transport and logistics | 115 | 16 | 6 | 81 |
| Automotive and assembly | 304 | 10 | 5 | 93 |
| Media and entertainment | 34 | 19 | 5 | 40 |
| Healthcare systems and services | 99 | 7 | 5 | 25 |
| Retail | 260 | 9 | 4 | 150 |
| Chemicals | 46 | 16 | 4 | 63 |
| Aerospace and defense | 25 | 10 | 3 | 23 |
| Consumer packaged goods | 124 | 15 | 3 | 121 |
| Basic materials | 101 | 17 | 3 | 122 |
| Electric power and natural gas | 67 | 23 | 3 | 104 |
| Infrastructure | 60 | 10 | 3 | 74 |
| **Total** | **3,059** | 17 | 5 | **1,850** |

[1] Global.

As seen in Exhibit 3, industries such as high tech and pharmaceuticals and medical products rank in the top quartile for most of these factors, while industries such as electric power and natural gas (EPNG) and infrastructure rank toward the lower quartile across many factors.

Exhibit 3

## Cloud value potential, as measured by EBITDA impact, varies by industry.

**EBITDA uplift enabled by cloud, by industry (2030 forecast)**
**Forbes 2000 largest companies by revenue**

□ Low ▨ Medium ■ High

| | 2030 revenue | IT spend intensity[1] | AI value from cloud[2] | IoT value from cloud[2] | Automation value from cloud[2] | 2030 revenue, $ billion | 2030 EBITDA, $ billion | Base EBITDA margin, % | 2030 EBITDA uplift from cloud, $ billion | **2030 EBITDA margin uplift from cloud,** % |
|---|---|---|---|---|---|---|---|---|---|---|
| High tech | ▨ | ■ | ■ | ▨ | ■ | 2,293 | 581 | 25 | 186 | **8** |
| Pharmaceuticals and medical products | □ | ▨ | ■ | ▨ | ▨ | 2,009 | 556 | 28 | 156 | **8** |
| Oil and gas | ■ | □ | ▨ | ■ | □ | 5,517 | 914 | 17 | 395 | **7** |
| Travel | □ | ▨ | ■ | □ | ■ | 205 | 49 | 24 | 14 | **7** |
| Telecommunications | ▨ | ■ | ■ | ▨ | □ | 3,000 | 897 | 30 | 211 | **7** |
| Banking | ▨ | ■ | ▨ | □ | ■ | 4,119 | 1,516 | 37 | 253 | **6** |
| Advanced electronics and semiconductors | ■ | ▨ | ▨ | ■ | ▨ | 6,237 | 994 | 16 | 356 | **6** |
| Insurance | ■ | ▨ | ■ | ▨ | □ | 4,511 | 455 | 10 | 256 | **6** |
| Transport and logistics | ▨ | ▨ | ▨ | ■ | ▨ | 2,080 | 327 | 16 | 115 | **6** |
| Automotive and assembly | ■ | □ | ▨ | ■ | ▨ | 5,573 | 567 | 10 | 304 | **5** |
| Media and entertainment | □ | ■ | ▨ | ▨ | ▨ | 685 | 131 | 19 | 34 | **5** |
| Healthcare systems and services | ▨ | ■ | □ | ■ | □ | 2,051 | 149 | 7 | 99 | **5** |
| Retail | ■ | ▨ | ▨ | ▨ | ▨ | 6,118 | 565 | 9 | 261 | **4** |
| Chemicals | □ | □ | ▨ | □ | ▨ | 1,211 | 199 | 16 | 46 | **4** |
| Aerospace and defense | □ | ▨ | □ | ▨ | ■ | 725 | 74 | 10 | 25 | **3** |
| Consumer packaged goods | ▨ | □ | ▨ | □ | ▨ | 3,743 | 554 | 15 | 124 | **3** |

Continued on next page

Exhibit 3, continued

**EBITDA uplift enabled by cloud, by industry (2030 forecast)**
**Forbes 2000 largest companies by revenue**

□ Low 🟦 Medium ⬛ High

| | 2030 revenue | IT spend intensity[1] | AI value from cloud[2] | IoT value from cloud[2] | Auto-mation value from cloud[2] | 2030 revenue, $ billion | 2030 EBITDA, $ billion | Base EBITDA margin, % | 2030 EBITDA uplift from cloud, $ billion | **2030 EBITDA margin uplift from cloud,** % |
|---|---|---|---|---|---|---|---|---|---|---|
| Basic materials | 🟦 | □ | □ | 🟦 | 🟦 | 3,227 | 561 | 17 | 101 | **3** |
| Electric power and natural gas | 🟦 | 🟦 | □ | 🟦 | □ | 2,278 | 526 | 23 | 67 | **3** |
| Infrastructure | 🟦 | 🟦 | □ | □ | ⬛ | 2,217 | 220 | 10 | 60 | **3** |

[1] As percent of revenue.
[2] As percent of uplift on revenue.

For example, pharma and medical products account for a relatively small revenue share among the Forbes 2000, with 75 companies generating $2 trillion of revenue out of a total of $58 trillion of revenues forecast for 2030. However, the IT intensity at 3.4 percent of revenue puts the pharma and medical products sector in the middle of the pack, and the AI potential is especially high. The capability to use cloud-based analytics to quickly discover and test new drugs and then market them appropriately, therefore, provides a distinct competitive advantage. Moderna, for example, developed its mRNA research-and-development platform on public cloud and was able to develop the first mRNA COVID-19 vaccine during the pandemic ahead of much larger peers.

In contrast, industries such as EPNG fall toward the lower end of the cloud value spectrum because they have middle-of-the road revenue share (104 companies generate $2.3 trillion of revenue) and IT intensity (2.6 percent of revenues), and lower impact from cloud-enabled analytics and automation use cases. These use cases do not fundamentally drive the profitability of the industry in the same way as use cases in pharma do. Instead, a utility's profitability is driven by the cost of capital investment in power plants, transmission, and distribution, with a regulated level of profit above cost recovery. This model makes it more challenging to drive outsized profit from cloud investments.

## Opportunities also vary by region

Asia has the highest cloud value potential, about $1.2 trillion, by 2030. While Asian companies lag American companies in their current levels of cloud adoption, they have the highest regional revenue share (38 percent) of Forbes Global 2000 companies. In addition, Asia has a huge presence in the oil and gas and banking industries, both of which exhibit potential for large EBITDA gains. As cloud service providers (CSPs) expand their footprint in Asia, these large companies will be able to achieve outsized value from cloud.

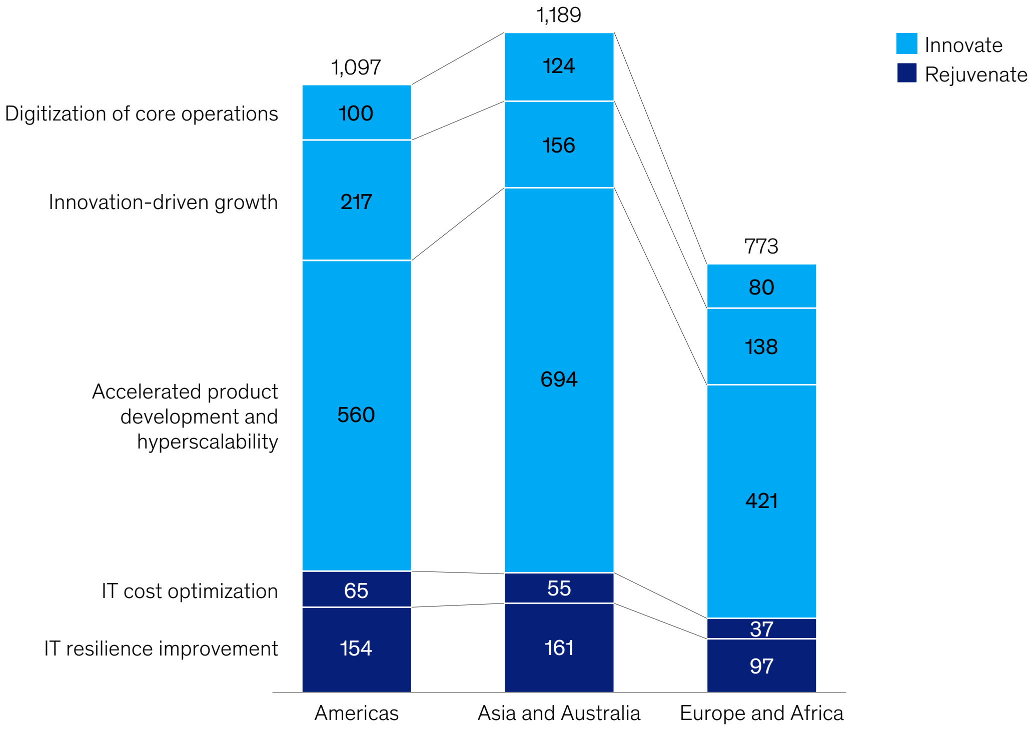
American institutions stand to capture about $1.1 trillion in cloud value. They make up 36 percent of the revenue of the Forbes Global 2000, and—given that three of the major CSPs started in North America—it is not surprising that cloud adoption there leads the rest of the world. North America's top industry driving cloud value is retail, which is expected to harvest nearly $161 billion in EBITDA gains from cloud by 2030, more than triple the value estimated for retail in the European Union and Asia.

EMEA has enormous potential in cloud as well, valued at $773 billion in EBITDA by 2030. This region has a relatively lower revenue share of the Forbes Global 2000 (25 percent), but its potential in cloud is buoyed by a favorable industry mix, with many top companies in high cloud-impact sectors such as automotive and assembly ($108 billion in EBITDA lift), and with more room to grow, given lower current adoption levels. Data sovereignty laws and regulatory pressures (for example, GDPR) may inhibit the migration and use of data that often drives significant cloud adoption, but there is a great deal of incremental value to be captured for companies that can navigate these forces (Exhibit 4).[5]

Exhibit 4

## Cloud's potential to capture value also varies by region.

**Estimated 2030 run-rate EBITDA impact,** $ billion



## Pioneers in each sector will use cloud to scale new technologies such as generative AI

We are seeing pioneers in many sectors, including more traditional ones such as banking, generate higher returns through pioneering approaches. For example, our research shows that incumbent banks will be increasingly challenged by growing competition from digital banks.

[5] "Projecting the global value of cloud," November 28, 2022.

Today, the difference in return on equity between a traditional bank and a digital bank is only 1 percent. In many cases, however, the digital business and its current return on equity doesn't reflect its growth potential. Within the next three to five years, we expect to see return on equity increases of 5 to 7 percent for digital banks. This is both a possibility and a threat—incumbent banks that are unwilling or unable to change and disrupt themselves will be commoditized or run out of business. Banks have a choice to either remain a commoditized business that strictly manages a balance sheet or evolve into a tech-enabled business that can compete in a new era of customer-centric financial experiences.

Looking ahead, we see eight tech trends noted in McKinsey's 2023 tech trends report that are most likely to drive outsized value in the next decade[6]:

- generative AI
- applied AI
- next-generation software development
- trust architectures and digital identity
- future of mobility
- immersive-reality technologies
- industrializing machine learning
- quantum technologies

The public cloud can help lower the barrier to entry into each of these tech trends by providing native services (to support quantum computing, for example), high scalability (which is needed to support generative AI's massive data sets), and accelerated development (for example, through next-gen software). As we have seen in the past ten years, digital natives can take advantage of these capabilities to gain outsized EBITDA value in the S&P 500, often at the expense of incumbents. This time around, incumbent businesses can also leverage cloud to gain their fair share of its potential value.

Let's take a closer look at one of these technologies. According to our research, generative AI is expected to contribute between $2.6 trillion and $4.4 trillion of value annually to the global economy.[7] The most-impacted sectors are likely to be high tech, retail, banking, travel and logistics, and advanced manufacturing, although generative AI will have broad impact beyond these sectors as well. Within functions, we expect generative AI to have the greatest productivity potential in marketing and sales, software engineering, and customer operations (Exhibit 5).

To achieve this value, organizations will need to train, tune, and deploy models based on the foundational models that underly generative AI use cases. Training generative AI foundation models requires expensive compute graphics processing units (GPUs) and becomes extremely capital-intensive at large scale (the cost to train ChatGPT4, for example, is more than $100 million). While few companies are expected to make that kind of investment, those that do will rely on public cloud due to its high scalability of compute resources and low up-front investment. Most other companies will access the foundation models and generative AI services that hyperscalers are investing in and developing.

For tuning and deployment (inferencing), we expect enterprises to take a hybrid approach. Model deployment and inferencing often involve embedding generative AI into real-time business processes that need to be close to existing systems and frontline staff. If these existing systems are in public cloud, as are many marketing platforms, for example, then the inferencing will most easily be done in public cloud. If the existing systems are on-premises, such as customer service or point-of-sale systems, then inferencing will more likely need to be done nearby. Similarly, the evolving regulation and privacy landscape will require businesses to think through what data they share and what they need to keep on-premises.

[6] "McKinsey Technology Trends Outlook 2023," McKinsey, July 20, 2023.

[7] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023. While it is clear that generative AI can drive big productivity gains, it remains to be seen exactly how much additional value could be attributed to cloud and added to our current estimate of $3 trillion EBITDA value of cloud by 2030.

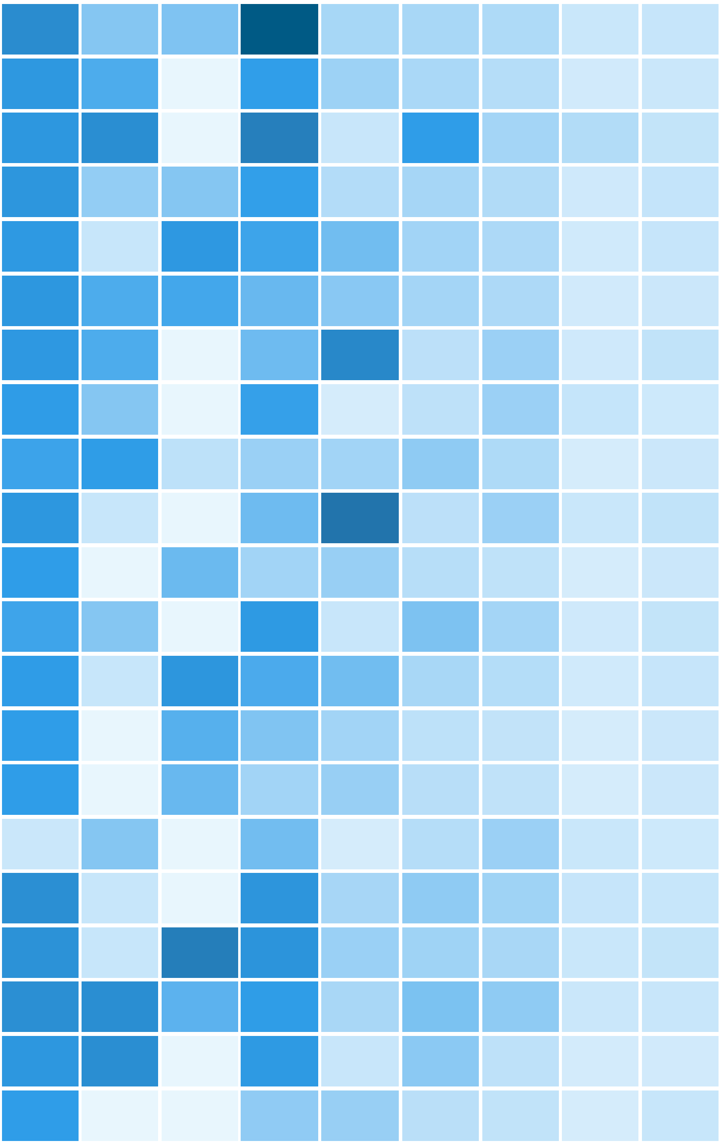
Exhibit 5

## Generative AI can have the greatest productivity impact on marketing and sales and software engineering.

**Generative AI productivity impact by business function**[1]

Low impact ▨ High impact

| | **Total,** % of industry revenue | **Total,** $ billion | Marketing and sales | Customer operations | Product and R&D | Software engineering | Supply chain and operations | Risk and legal | Strategy and finance | Corporate IT[4] | Talent and organization |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Total** | 1.3–2.1 | 2,600–4,400 | 760–1,200 | 340–470 | 230–420 | 580–1,200 | 280–530 | 180–260 | 120–260 | 40–50 | 60–90 |
| High tech | 4.8–9.3 | 240–460 | | | | | | | | | |
| Retail[2] | 1.2–1.9 | 240–390 | | | | | | | | | |
| Banking | 2.8–4.7 | 200–340 | | | | | | | | | |
| Travel, transport, and logistics | 1.2–2.0 | 180–300 | | | | | | | | | |
| Advanced manufacturing[3] | 1.4–2.4 | 170–290 | | | | | | | | | |
| Consumer packaged goods | 1.4–2.3 | 160–270 | | | | | | | | | |
| Healthcare | 1.8–3.2 | 150–260 | | | | | | | | | |
| Administrative and professional services | 0.9–1.4 | 150–250 | | | | | | | | | |
| Energy | 1.0–1.6 | 150–240 | | | | | | | | | |
| Education | 2.2–4.0 | 120–230 | | | | | | | | | |
| Basic materials | 0.7–1.2 | 120–200 | | | | | | | | | |
| Real estate | 1.0–1.7 | 110–180 | | | | | | | | | |
| Advanced electronics and semiconductors | 1.3–2.3 | 100–170 | | | | | | | | | |
| Chemicals | 0.8–1.3 | 80–140 | | | | | | | | | |
| Construction | 0.7–1.2 | 90–150 | | | | | | | | | |
| Public and social sector | 0.5–0.9 | 70–110 | | | | | | | | | |
| Media and entertainment | 1.5–2.6 | 60–110 | | | | | | | | | |
| Pharmaceuticals and medical products | 2.6–4.5 | 60–110 | | | | | | | | | |
| Telecommunications | 2.3–3.7 | 60–100 | | | | | | | | | |
| Insurance | 1.8–2.8 | 50–70 | | | | | | | | | |
| Agriculture | 0.6–1.0 | 40–70 | | | | | | | | | |



Note: Figures may not sum to totals, because of rounding.
[1] Excludes implementation costs (eg, training, licenses).
[2] Includes auto retail.
[3] Includes aerospace, defense, and auto manufacturing.
[4] Excludes software engineering.
Source: CIS/IHS Markit; "The economic potential of generative AI: The next productivity frontier," McKinsey, June 24, 2023; McKinsey Manufacturing and Supply Chain 360 assessment; McKinsey Sales Navigator; Oxford economics; internal experts and databases: McKinsey corporate business functions

It is fairly clear in these early days of generative AI that the technology hardware and software players that have invested early will stand to capture the most value from it. However, as its underlying services, models, and capabilities become more available and commoditized—we are already seeing a proliferation of open-source models—there will be significant opportunities for pioneering enterprises to create new revenue streams by combining generative AI offerings with their proprietary data and existing competitive advantages.

## What's impeding companies from realizing their share of cloud's potential value?

To deepen our understanding of how organizations are or are not capturing cloud value, we profiled cloud programs at more than 90 large enterprises. In all cases, we spoke live with senior executives—typically the CIO, CTO, head of infrastructure or cloud-program lead—who have insight and responsibility across the enterprise (Exhibit 6).

In our interviews, we began by taking stock of the basics, asking how far along these companies are in their cloud journeys, and how far they aspire to go. Companies in North America aspire to have more than half of their applications running in cloud (either via SaaS or public cloud) within three to five years (Exhibit 7).

Many companies are a long way from fulfilling their ambitions. With a few notable exceptions, most of the companies we profiled have achieved only a fraction of their aspirations for cloud adoption. Even including SaaS adoption, only 39 percent of companies have more than 30 percent of their applications running on public cloud.

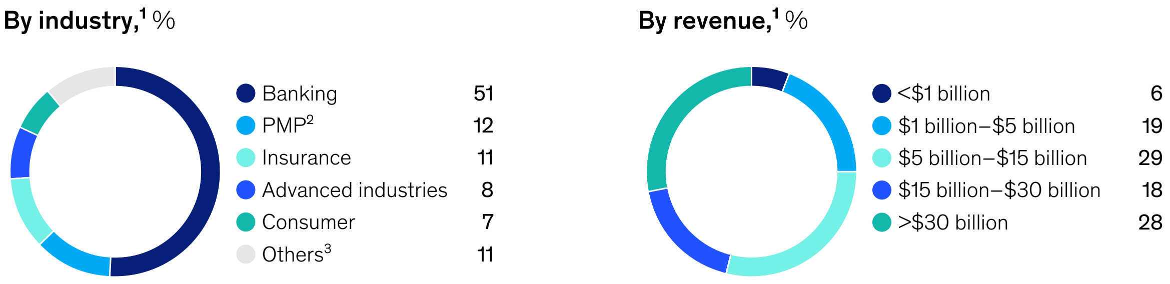
Why is it that so many companies still have so far to go? We might expect that the organizations with the longest-running cloud programs would have made the most progress, but that doesn't seem to be the case. Several companies we profiled have been running cloud programs since 2015 but still have fewer than 20 percent of their applications in public cloud. Meanwhile, some companies whose cloud programs have been running half as long have more than a 60 percent cloud adoption rate (Exhibit 8).[8] The lesson is that a cloud program can't be put on autopilot; it has to be carefully tended and built up over time. Companies that move with purpose reach high levels of cloud adoption quickly.

[8] Interviews with cloud leaders at about 60 large organizations.

Exhibit 6

### We profiled more than 90 organizations to understand their progress on capturing value in cloud.

**Company demographics (n = 91)**

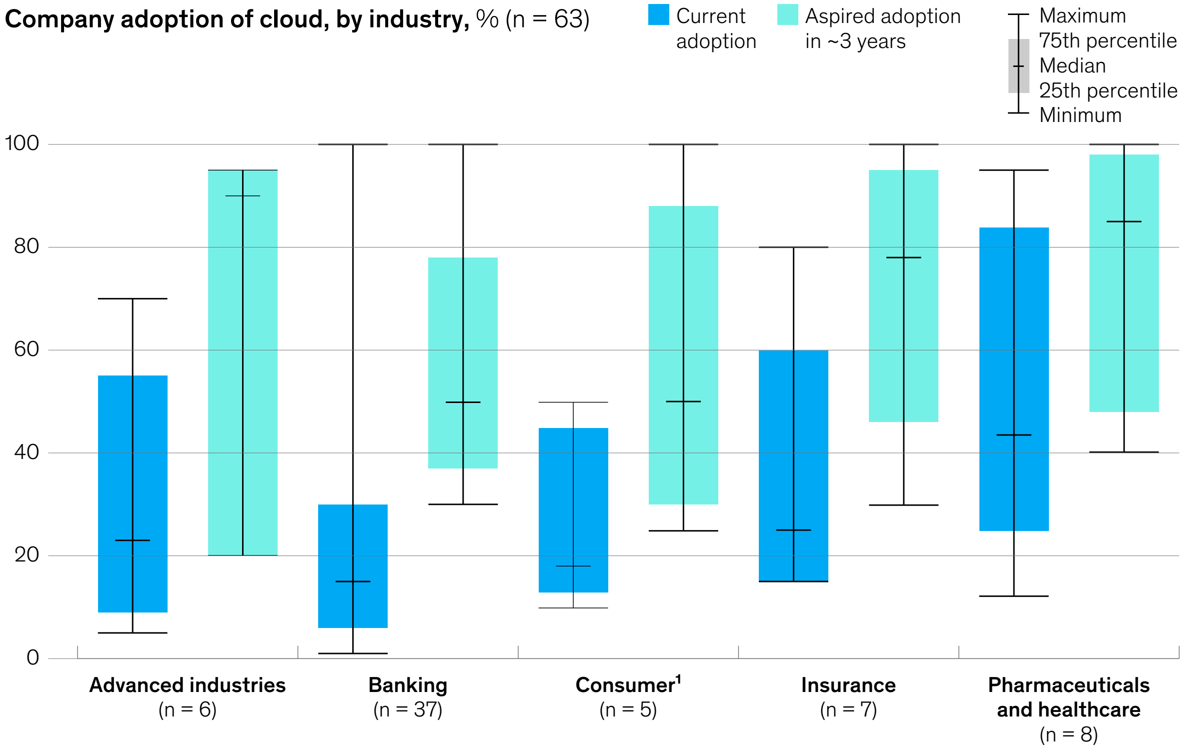


[1] North American and European organizations.
[2] Pharmaceuticals and medical products.
[2] Professional services, agriculture, airlines, R&D.

Exhibit 7

## Most organizations aspire to host more than 50 percent of their applications on cloud platforms in three years.

**Company adoption of cloud, by industry,** % (n = 63)



[1]Outlier removed due to small sample size.
Source: CloudSights

But even assuming an organization is successful in swiftly moving a significant portion of its applications to public cloud, there is still the question of value. Only about 10 percent of the companies we surveyed believe they have fully captured value at scale. Another roughly 50 percent have begun to see value capture in pockets. Often they have lowered IT costs by exiting data centers and are seeing pockets of value in one business domain or function but have yet to scale broadly across the organization. And the remainder? Forty percent of companies profiled have moved very few applications to cloud and therefore have seen little cloud value (Exhibit 9).[9]

Why is capturing value so hard? The main reason is that many tech leaders today still view cloud primarily as a successor to the series of hosting innovations like commodity x86 architectures, open-source Linux, and virtual machines running on a single server, all of which transformed the cost structure of application hosting. But cloud is much more than a hosting innovation, and its value only comes from making the changes to the business model and operating model that enable cloud's true advantages.

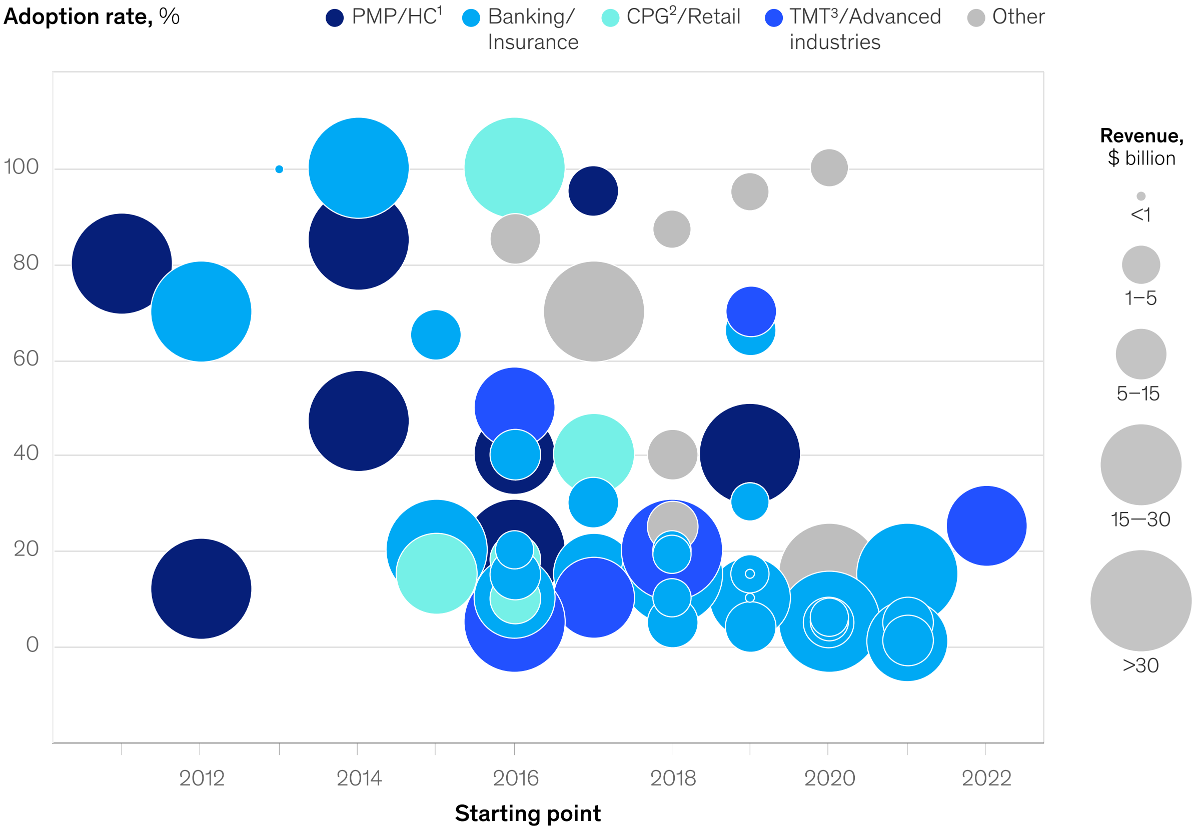
To reap cloud's full value, tech leaders need to address three significant sources of complexity. The first is business alignment. Companies often struggle to align the executive team around the link between cloud investments and business priorities and outcomes. An underlying problem is that IT makes much of the cloud

[9] Interviews with cloud leaders at about 60 large organizations.

Exhibit 8

## Many companies have low cloud adoption (less than 20 percent) despite being years into their cloud programs.

**Cloud-journey start date versus % of applications in cloud,** n = 61



[1] Pharmaceuticals and medical products; healthcare.
[2] Consumer packaged goods.
[3] Technology, media, and telecommunications.
Source: CloudSights

investment, while the benefits are often seen in other parts of the organization. This creates disincentives for the IT organization to make sufficient investments. It is therefore critical both to demonstrate in clear dollars-and-cents terms the value of cloud and to align executives around the need to provide sufficient investment to deliver that value.

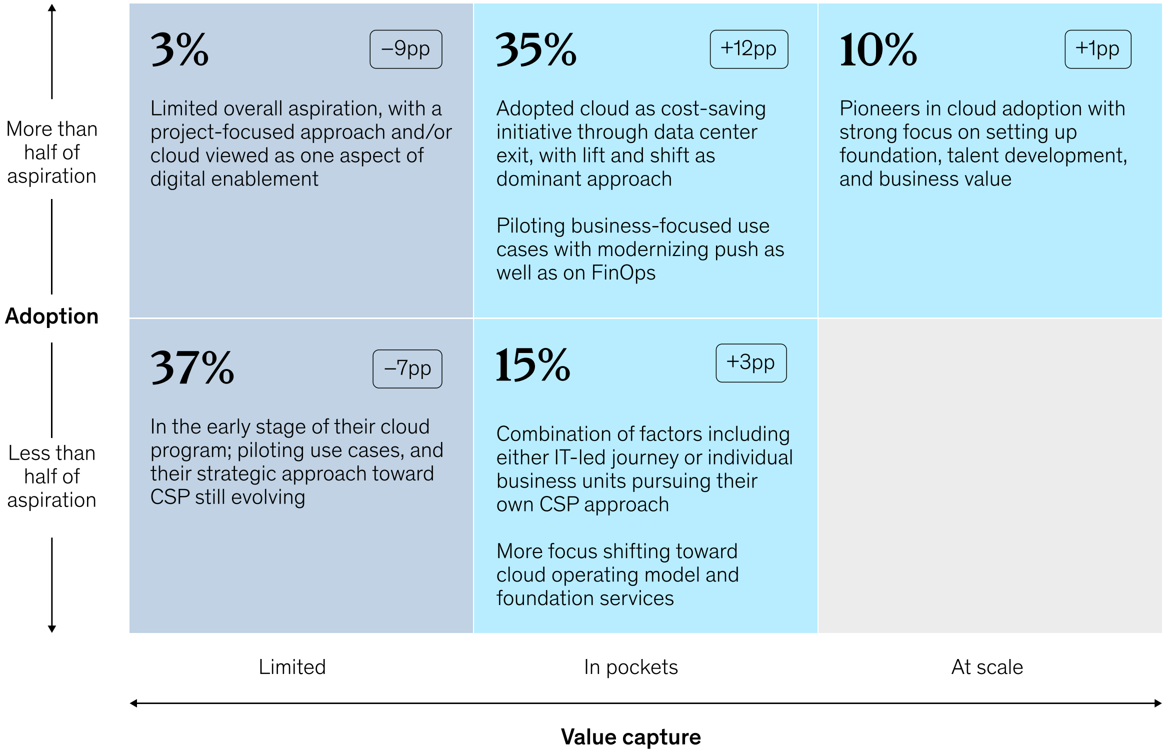
The second is architectural complexity. In large organizations, many existing applications have to be remediated or re-architected to run efficiently, securely, and resiliently in cloud. Without those changes, companies have found that some applications cost more to run in cloud. Remediation costs are often much more than companies are comfortable with, especially those that have already aggressively optimized their on-premises infrastructure. The economics of cloud adoption become much more attractive for companies when they focus on remediating applications that are critical to high-value use cases, using the migration process itself as a mechanism to force infrastructure optimization and avoid large capital investments in data centers.

Exhibit 9

## While 50 percent of companies are realizing value in pockets, only 10 percent are finding it at scale.

**Cloud value capture versus adoption,** % of organizations (n = 59)

□ Change from 2022



| Adoption | Limited | In pockets | At scale |
|---|---|---|---|
| More than half of aspiration | 3% (−9pp)<br>Limited overall aspiration, with a project-focused approach and/or cloud viewed as one aspect of digital enablement | 35% (+12pp)<br>Adopted cloud as cost-saving initiative through data center exit, with lift and shift as dominant approach<br>Piloting business-focused use cases with modernizing push as well as on FinOps | 10% (+1pp)<br>Pioneers in cloud adoption with strong focus on setting up foundation, talent development, and business value |
| Less than half of aspiration | 37% (−7pp)<br>In the early stage of their cloud program; piloting use cases, and their strategic approach toward CSP still evolving | 15% (+3pp)<br>Combination of factors including either IT-led journey or individual business units pursuing their own CSP approach<br>More focus shifting toward cloud operating model and foundation services | |

Source: CloudSights

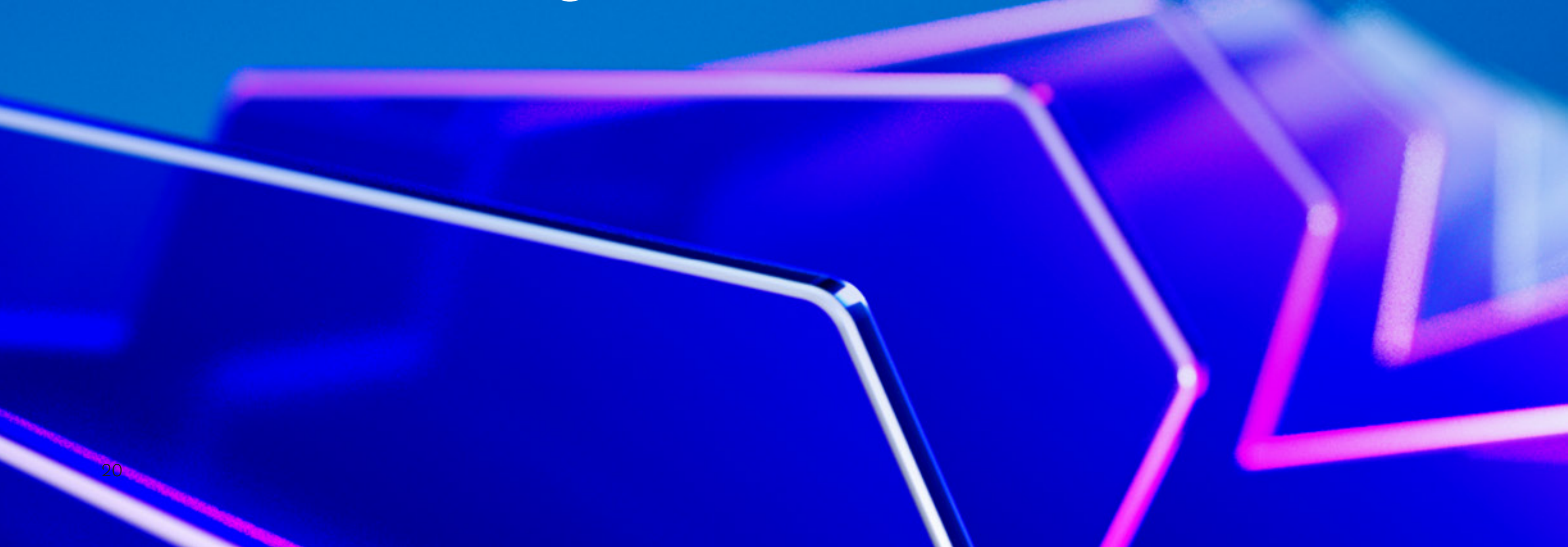
The third is organizational complexity. Cloud adoption requires coordinated action across IT *and* the business. The infrastructure team needs to build and enable the cloud foundation, the application teams need to migrate or remediate their applications, and the business teams need to translate the newfound technology capabilities into business benefits. Both teams need to work in a highly coordinated way, with joint objectives, to avoid sprawling initiatives, duplicated capabilities, and delayed migration. Overcoming this organizational complexity requires an operating model that includes a cross-functional business and IT cloud-governance structure, team structures built around products, strong engineering talent, and a program to drive, manage, and reinforce the changes.

We will explore the solutions to these issues in greater detail in Part III.

## Part II

# How can my company maximize its cloud ROI?

Breakevens, value leakage, and generative AI's role



**As any business leader knows,** it takes investment to capture value. What's missing in many cloud conversations, however, is a clear understanding of how much investment is needed, what kind of returns can be expected on those investments, and when the returns can be expected to kick in.

We have developed a tool—CloudSights: ROI Engine—to enable a clearer understanding of cloud's ROI (see sidebar "About CloudSights: ROI Engine"). Our analysis revealed that the ROI for a typical cloud program focused on a single business domain can be as high as 180 percent, although it varies with the number of domains, the level of investment appetite, and the investment time horizon. Despite the promising ROI potential, companies still suffer from misconceptions and lack of clarity when making their individual ROI calculations. That's a cloud killer. Unless companies can calculate their cloud ROI curve with sufficient rigor, it's hard to get the necessary commitment from the CEO or CFO to align the organization around what it will take to capture cloud's value.

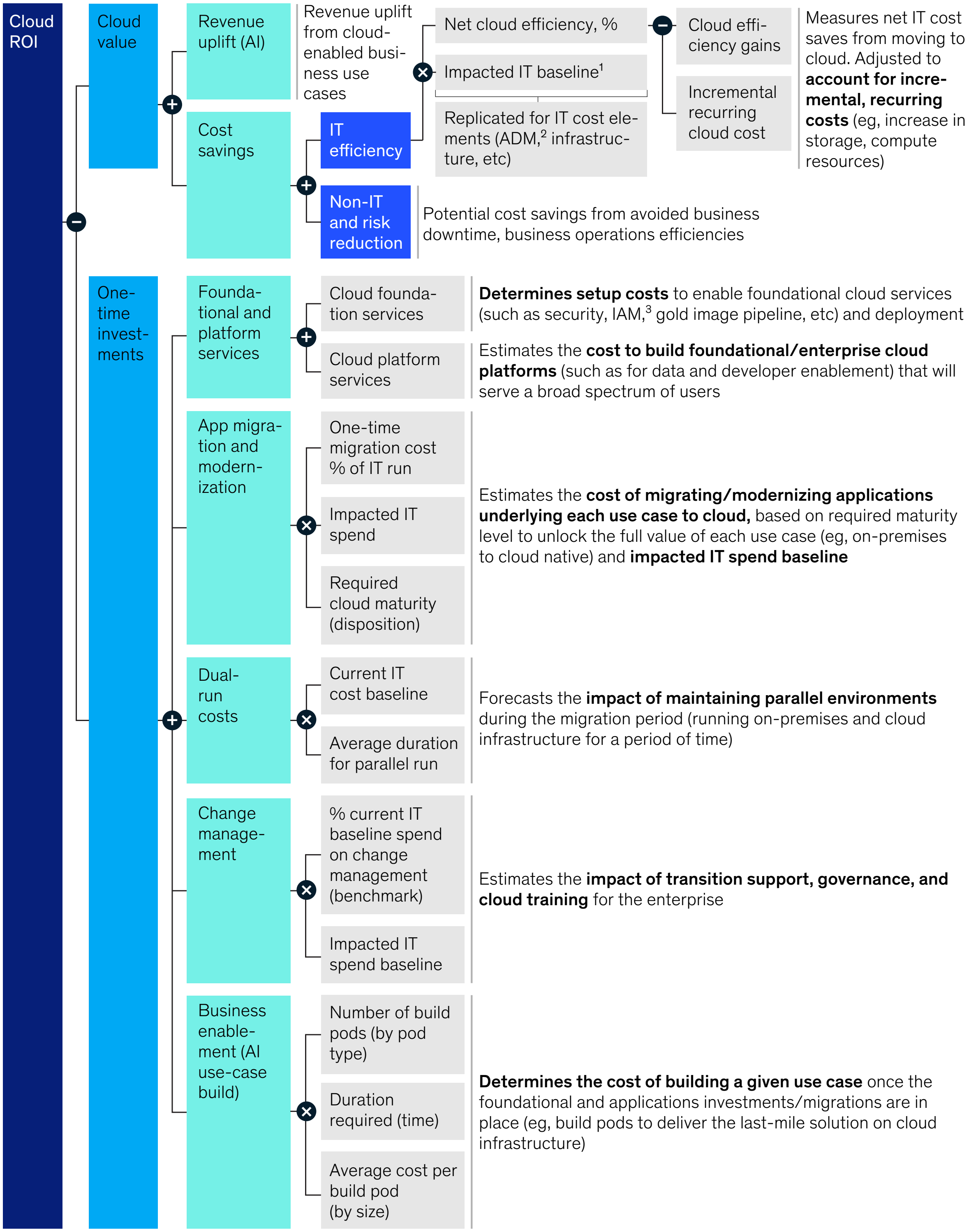
An important starting point is to identify the drivers of investment cost and sources of cloud value in sufficient detail to measure them (Exhibit 10). We articulated the benefits side of the equation in Part I, including the IT cost efficiencies as well as the non-IT business value and risk reduction benefits.

On the investment side, factors include the setup costs to enable foundational cloud services, the costs of modernizing applications and/or migrating them to cloud, of maintaining parallel on-premises and cloud environments during the transition period, and of change management (including transition support, governance, and training). Many of these calculations need to be made at the workload level so it's clear which workloads need to be migrated to cloud, the cost of each one, and how many need to be migrated before the business breaks even.

Exhibit 10

## Accurately calculating cloud ROI requires understanding the sources of cloud value and the drivers of investment costs.

**Not exhaustive**



Cloud ROI = Cloud value − One-time investments

Cloud value = Revenue uplift (AI) + Cost savings

Revenue uplift (AI): Revenue uplift from cloud-enabled business use cases

Cost savings = IT efficiency + Non-IT and risk reduction

IT efficiency = Net cloud efficiency, % × Impacted IT baseline[1]; Replicated for IT cost elements (ADM,[2] infrastructure, etc)

Net cloud efficiency, % = Cloud efficiency gains − Incremental recurring cloud cost

Measures net IT cost saves from moving to cloud. Adjusted to **account for incremental, recurring costs** (eg, increase in storage, compute resources)

Non-IT and risk reduction: Potential cost savings from avoided business downtime, business operations efficiencies

One-time investments = Foundational and platform services + App migration and modernization + Dual-run costs + Change management + Business enablement (AI use-case build)

Foundational and platform services = Cloud foundation services + Cloud platform services

- Cloud foundation services: **Determines setup costs** to enable foundational cloud services (such as security, IAM,[3] gold image pipeline, etc) and deployment
- Cloud platform services: Estimates the **cost to build foundational/enterprise cloud platforms** (such as for data and developer enablement) that will serve a broad spectrum of users

App migration and modernization = One-time migration cost % of IT run × Impacted IT spend × Required cloud maturity (disposition)

Estimates the **cost of migrating/modernizing applications underlying each use case to cloud,** based on required maturity level to unlock the full value of each use case (eg, on-premises to cloud native) and **impacted IT spend baseline**

Dual-run costs = Current IT cost baseline × Average duration for parallel run

Forecasts the **impact of maintaining parallel environments** during the migration period (running on-premises and cloud infrastructure for a period of time)

Change management = % current IT baseline spend on change management (benchmark) × Impacted IT spend baseline

Estimates the **impact of transition support, governance, and cloud training** for the enterprise

Business enablement (AI use-case build) = Number of build pods (by pod type) × Duration required (time) × Average cost per build pod (by size)

**Determines the cost of building a given use case** once the foundational and applications investments/migrations are in place (eg, build pods to deliver the last-mile solution on cloud infrastructure)

[1] Requires IT costs to be estimated at various levels (eg, certain areas may be applicable at function or domain level vs use case).
[2] Application delivery management.
[3] Identity and access management.

## About CloudSights: ROI Engine

CloudSights: ROI Engine is a tool powered by a model that weighs the benefits of cloud adoption against the investments required to achieve them. It is designed to optimize cloud investment, decision making, and prioritization around use cases that are unlocked or accelerated by cloud. The ROI Engine considers numerous variables to assess the outlook for organizations based on their industry, size, and level of cloud maturity, among other factors. The underlying model that powers the ROI Engine's analysis is based on data collected via McKinsey Global Institute, firm leaders, third-party data sources, and client experience.

## Three typical cloud ROI paths to consider

One of the most pressing questions that cloud leaders face today is how swift and ambitious they should be in terms of their cloud adoption and business transformation. There are a variety of different paths to satisfy an organization's investment appetites and business priorities, each of which has an effect on its ROI curve.

For a closer understanding of ROI curves, let's consider three scenarios:

— **An IT-led transformation** focuses on improving the efficiency of an organization's technology *and* reducing the risk of downtime or security breaches. It typically consists of reducing the on-premises footprint by exiting data centers and migrating to cloud, building a resilient and secure cloud foundation, and modernizing where necessary in cloud.

— **A single-domain-focused transformation** builds a cloud capability around one business area, such as a customer journey or a complete process, by targeting investments in digital innovation enabled by cloud platform services.

— **A cross-domain transformation**, the most aggressive approach, aims to transform the whole business by building cloud capabilities across multiple business domains.

Each path requires different levels of investment, has a different timeline in terms of how quickly those investments will pay off, and has a different long-term ROI. To better understand how the ROI curve will change for each of these archetypes, let's look at a hypothetical Forbes 2000 pharma company with approximately $19 billion in revenue and $650 million in IT costs (Exhibit 11). If this pharma company chooses the IT-led transformation, a well-executed program will take six years to break even and eight years to achieve 145 percent ROI. The aggressive, cross-domain transformation, on the other hand, potentially breaks even within just three years and achieves a 350 percent ROI over the same eight-year period, but it requires much more investment up front.

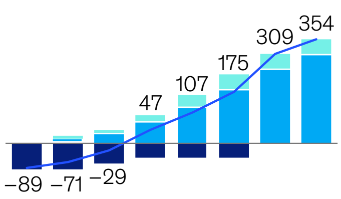
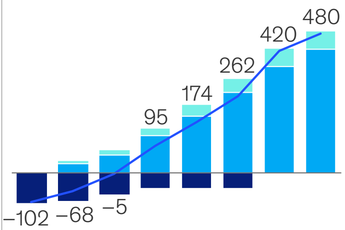
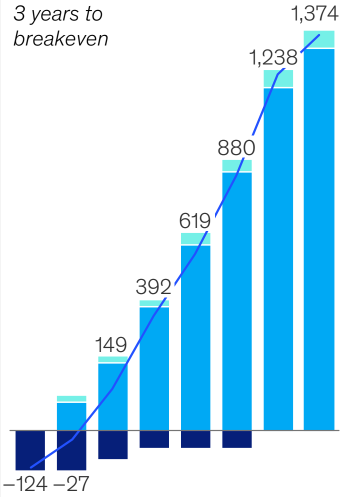
The reason for the better outcomes from the more-aggressive pathway is that many of the initial costs are one-off investments, such as building out the cloud foundation. These investments are amortized quickly because so many workloads are migrated using the same infrastructure and services. This approach has the biggest payoff but also the highest risk, requiring companies to be comfortable with significant up-front investments and confident in their capabilities to deliver the work.

Exhibit 11

## ROI outcomes are strongly influenced by a company's approach to transformation.

**Transformation scenarios' ROI for average Forbes Global 2000 pharma company with ~$18.8 billion in revenue and $650 million in IT spend**

▨ IT efficiency ■ Business benefits[1] ■ Investments — Net for year

| | **A. IT-led transformation** | **B. Cloud-enabled R&D transformation**<br>*(Deep dive follows)* | **C. Cloud-enabled customer- and partner-experience transformation** |
|---|---|---|---|
| **8–year ROI,** % | **145** | **175** | **350** |
| **EBITDA gain and investment forecast,**<br>$ million | *6 years to breakeven*<br><br> | *4 years to breakeven*<br><br> | *3 years to breakeven*<br><br> |
| **New recurring cloud costs,**[2]<br>$ million | 2 4 6 9 12 17 19 21 | 4 8 14 19 25 32 35 39 | 9 21 36 57 83 114 130 147 |
| Year: | 1 2 3 4 5 6 7 8 | 1 2 3 4 5 6 7 8 | 1 2 3 4 5 6 7 8 |
| **What is included?** | • Core-systems migration<br>• Tech debt/legacy reduction<br>• Corporate back-office analytics and automation | • Core eClinical SaaS platforms (eg, EDC, CTMS, eTFM)<br>• RWD[3] platform build<br>• Next-generation capabilities (eg, DCT[4], telehealth) | • Ad and marketing low-code and no-code (LCNC) platform(s)<br>• Ecosystem integration (with, eg, RWD[3] platform approach, CRO[5] and partner data)<br>• Digital-experience drivers for improved regulatory engagement |

[1] Includes business benefits from IT resiliency improvement, cloud-accelerated / cloud-unlocked digital use cases.
[2] Incremental recurring cloud spend driven by delivering new digital use cases on cloud.
[3] Real-world data.
[4] Decentralized clinical trials.
[5] Clinical research organization.
Source: External research reports (IDC, IHS Markets), Forbes Global 2000 company financials, McKinsey cloud transformation examples (benchmarks)

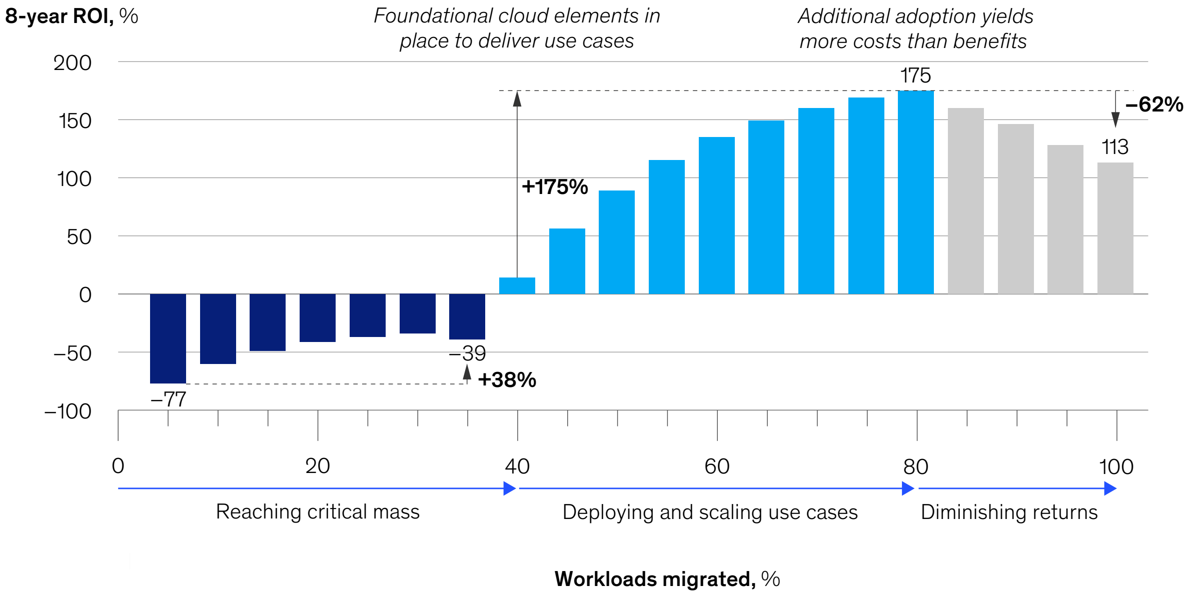
No matter which pathway a business chooses, it will have two important inflection points: the breakeven moment, when the program's ROI turns positive, and the point of diminishing returns from modernizing or migrating the remaining workloads. These are the workloads that don't generate significant value for the business. Being clear about each of these inflection points is critical. Understanding the breakeven point sets the right expectations for business leadership and can help maintain the resource commitment. Understanding where diminishing returns kick in can help determine when to start dialing back or reallocating investments.

If the pharma company described above chooses to pursue the middle path, a single-domain transformation, breakeven comes only after migrating 38 percent of workloads to cloud. Significant benefits then continue to accrue until the company has migrated approximately 80 percent of its workloads, which marks the point of diminishing returns (Exhibit 12).

Exhibit 12

## Peak ROI of 175 percent is achieved at 80 percent cloud adoption.

**ROI profile vs adoption: Single-domain transformation for average Forbes Global 2000 pharma company with ~$18.8 billion in revenue**



Cloud programs usually break even when they meet two important benchmarks, generally within three to four years. The first is when core cloud foundations (security, app patterns, landing zones, and so on) and cloud platform services (developer and data enablement) are complete. The cost of these investments is typically about 15 to 30 percent of an organization's baseline, pre-cloud IT spend. The second is when workloads on cloud reach a critical mass of typically around 30 to 40 percent of applications. The cost of this migration can be around 10 to 15 percent of an organization's baseline pre-cloud IT spend.

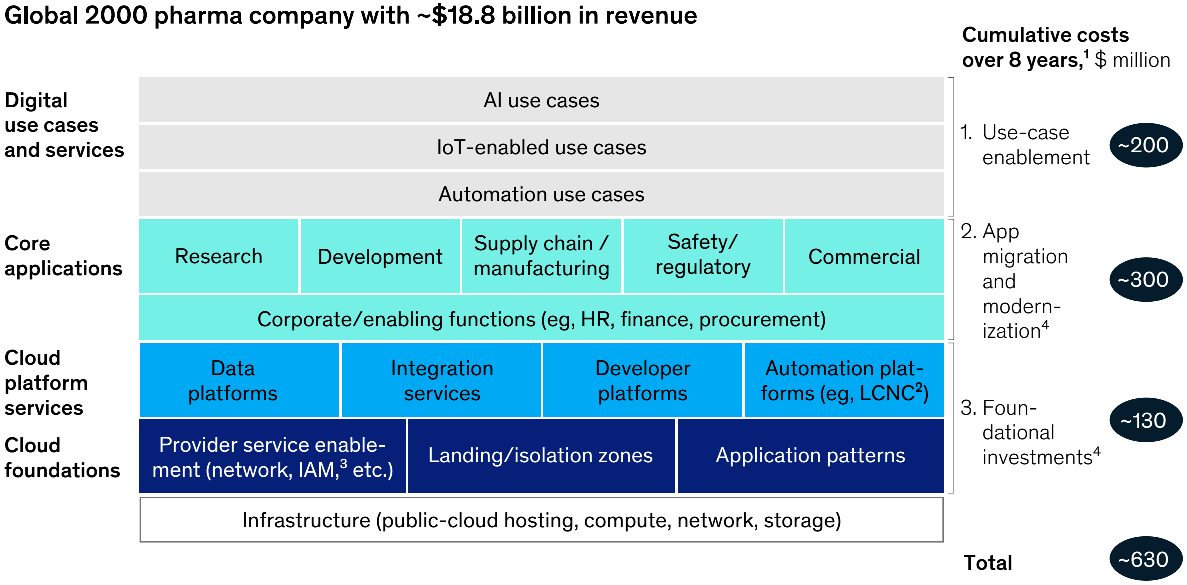
Three to four years to break even may seem daunting, especially in a somewhat uncertain economic environment. To build confidence within the organization, some cloud leaders may choose to start with more-limited cloud experiments, focused perhaps on a single use case or just a handful of applications. This approach can help generate some early wins. But transformational value is only possible with more significant commitments.

Each of the three pathways will call for three main kinds of investment: foundational investments (to establish strong core capabilities before beginning at-scale transformations), application migration and modernization (so applications can take advantage of cloud's benefits), and use-case enablement (to build value). Each path will distribute these investments differently. In the case of the pharma company that has chosen the intermediate, single-domain play, about 20 percent of its total $630 million investment will fund foundational investments, nearly half will go to app migration and modernization, and just over 30 percent will go to use-case enablement (Exhibit 13).

Exhibit 13

## Transformation of even a single business domain requires investment in three key areas.

**Illustrative cloud-based enterprise architecture: Single-domain transformation for average Forbes Global 2000 pharma company with ~$18.8 billion in revenue**



Cumulative costs over 8 years,[1] $ million
Digital use cases and services
AI use cases
IoT-enabled use cases
Automation use cases
Core applications
Research
Development
Supply chain / manufacturing
Safety/ regulatory
Commercial
Corporate/enabling functions (eg, HR, finance, procurement)
Cloud platform services
Data platforms
Integration services
Developer platforms
Automation platforms (eg, LCNC[2])
Cloud foundations
Provider service enablement (network, IAM,[3] etc.)
Landing/isolation zones
Application patterns
Infrastructure (public-cloud hosting, compute, network, storage)
1. Use-case enablement
~200
2. App migration and modernization[4]
~300
3. Foundational investments[4]
~130
Total
~630

[1] Inclusive of both one-time and recurring costs. Estimates are aggregate cost from 2023–30 (starting adoption ~30%, target adoption 80%, in terms of workloads on cloud).
[2] Low-code and no-code development platforms.
[3] Identity and access management.
[4] Change management (transition support) cost included in cloud foundation costs, while dual-run cost of duplicative computing environments is included under app migration cost.
Source: McKinsey client examples

## Barriers to maximum ROI

There are three main factors that can completely eliminate many potential cloud benefits and leave programs with a low or negative ROI (Exhibit 14).

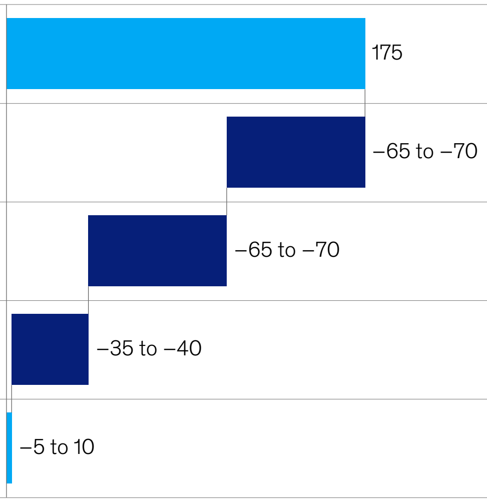
### Unrealized use cases

Sometimes cloud programs will make all the right investments in foundations and app migration and modernization but will fail to implement the most valuable use cases. A common reason is that organizations are so focused on the immediate reward of IT cost savings that they neglect the longer-term (but much larger) benefits that come from the cloud-enabled deployment of new digital use cases. Companies in effect settle for a much smaller slice of value by not taking full advantage of their cloud investments.

Exhibit 14

## Three mistakes drive real-life cloud programs’ failure to achieve optimal returns.

**Value-leakage drivers example: Single-domain transformation for average Forbes Global 2000 pharma company with ~$18.8 billion in revenue**



| | **8-year ROI (2023–30),** % | **Description** |
|---|---|---|
| Potential cloud ROI | 175 | Estimated maximum ROI (optimal prioritization, sequencing, execution) |
| Unrealized use cases | –65 to –70 | Focusing on IT cost savings vs rapid experimentation and deployment of new digital capabilities on cloud |
| Cloud sprawl | –65 to –70 | Building duplicative cloud foundations and platforms (eg, siloed by business area, multi-CSP) |
| Stalled adoption | –35 to –40 | Stalling at ~50% adoption on cloud, resulting in high stranded costs (tech debt, data center exit) and reduced reusability of analytics and data |
| Reality of today’s cloud programs | –5 to 10 | Where many organizations end up, despite being years into their cloud journey |

In the pharmaceutical company analysis, for example, 20 percent of the one-time cost is in establishing the cloud foundation itself. Once this investment has been made, every extra dollar of investment to enable use cases, such as building cloud data pipelines and AI models, has a much larger incremental ROI.

### Cloud sprawl

Without sufficient care, app migrations can lead to a complicated, unwieldy cloud architecture, which we call “cloud sprawl.” Companies build duplicative cloud foundations and platforms, often siloed by business area, or end up with complex, inefficient multi-CSP designs.

This often occurs when cloud leaders take an overly democratic approach to implementation. During a cloud transition, they are so focused on weighing various stakeholders’ needs and desires that they prioritize these concerns over the need to build an efficient, streamlined cloud program. As a result, they end up with overly complicated, inefficient cloud architectures that cost far more than they should. It is important to understand a broad range of concerns and respond to those that are crucial. But when undertaking something as ambitious as transitioning to cloud, tech leaders need to be decisive about developing an architecture that best serves overall business interests.

Eight years ago, a major financial services company began adopting cloud, initially giving flexibility to the IT teams and business units, and dabbling with a multicloud approach. But the company soon realized that having workloads in both public and private cloud environments competing for workloads was creating unnecessary complexity and friction. It also found that trying to be flexible or cloud agnostic slowed down adoption. In the last two to three years, the company has shifted to building a mature cloud services platform, enabling developers to build, deploy, and operate applications and infrastructure while maintaining focus on security. This has led to many benefits, including a 30 percent increase in developer productivity.

**Stalled adoption**

In the early phase of a cloud program, executives tend to become concerned about costs, resource commitments, and slow returns—especially if they have an incomplete understanding of how long it takes for cloud investments to pay off. Some companies may choose to pull the plug entirely. Others will choose half-measures, slowing investments when they should be accelerating, or looking for exits when they should be doubling down. They are looking to hedge their bets, but they may hedge themselves out of the game, and their cloud program can stall out.

This sort of hesitancy can significantly raise the cost of a cloud transition. One of the most common ways that this happens is that organizations decide to also keep the on-premises system for a while, clinging to the notion of "going back" if difficulties arise. We estimate this dual-run cost of maintaining a parallel environment for six months after migrating applications at 5 to 7 percent of a company's baseline IT spend. Some organizations maintain the parallel environments for much longer—as much as 18 months for a given application—or, in the worst case, simply maintain dual on-premises and cloud environments indefinitely. The costs can quickly become a massive burden.

## How generative AI could reduce cloud value leakage and increase ROI

We are still in the early days of the generative AI revolution, but it's already clear that generative AI's impact on cloud could be significant in terms of improving the ROI of cloud programs. Major hyperscalers are rolling out new generative AI–enabled services that can help companies improve the economics of some use cases and open access to new ones. How well companies take advantage of these services depends on a host of reasons, including their cloud maturity and the strength of their cloud foundations. Foundation models that rely on customer data that is already in cloud, for example, can be more easily and cheaply trained and then deployed using cloud-based generative AI services.

The other source of value is the use of generative AI tools to reduce the costs of application migration and remediation. Our analysis has shown that these tools could improve the ROI of cloud programs by 75 to 110 percentage points (Exhibit 15). This improvement is driven by the reduction in one-time application migration and remediation costs, accelerated migration timelines, and increased developer productivity for developing new features in cloud with generative AI. For example, early experiments applying a generative AI–enabled approach to mainframe remediation and migration have demonstrated a more than 40 percent reduction in the cost and time it takes to migrate the application. Similarly, experiments with generative AI for new feature development have demonstrated a 30 to 50 percent increase in productivity in prototyping and deploying new code.

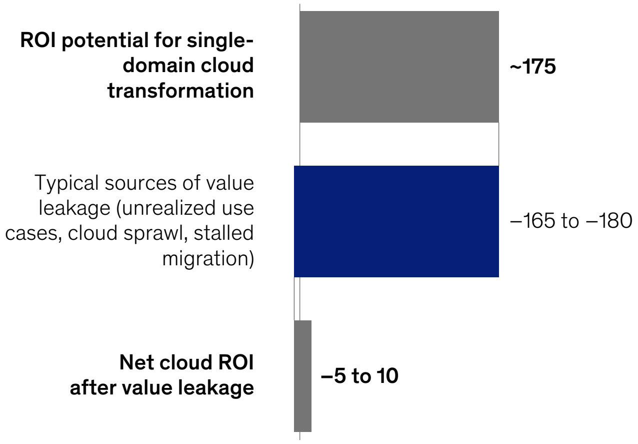
Importantly, these experiments show that generative AI is not just useful for code conversion and generation but can be applied to augment humans across the end-to-end process of application modernization, which can be broken down into three major steps: discovery and assessment, planning and design, and conversion. In the discovery and assessment phase, generative AI tools can parse millions of lines of outdated code and translate them into plain English so experts can understand which code blocks drive which functions. In the planning phase, generative AI tools can help map out and prioritize which code blocks to modernize and to what target state, including whether to add new capabilities. Finally, in the conversion phase, generative AI tools can translate the legacy code into plain English, generate new code in the target language, and automatically create test scripts to ensure that the new code provides the same or better output compared to the legacy code.

Exhibit 15

## The next wave of generative AI tools could help plug leaks in value and increase cloud program ROI by 75 to 110 percentage points.

### Example single-domain-focused transformation for a Forbes 2000 pharma company (estimated over a 7-year period)

**Today's reality: Value leakage in typical cloud transformation,** net ROI, %



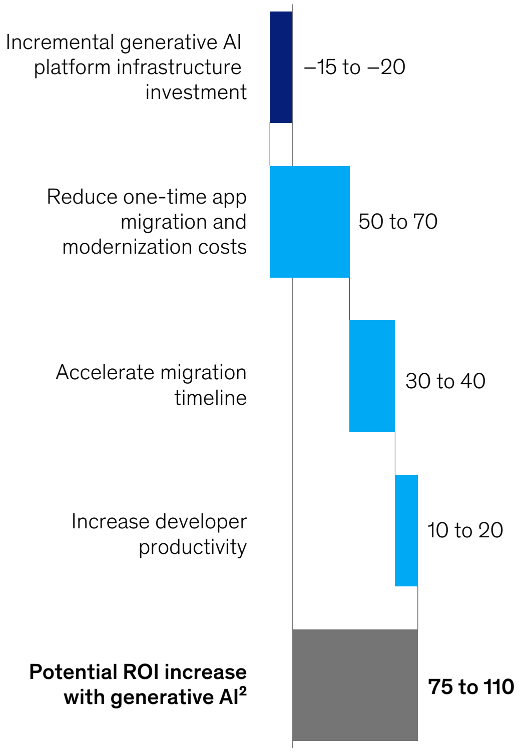
**Description**

**Potential ROI for single-business-domain transformation** (including migrating from 30% current workloads on public cloud to 80% by 2030)

**Cloud program value leakage driven by three factors:** unrealized use cases (–65% to –70% ROI), cloud sprawl (–65% to –70% ROI), and stalled migration (–35% to –40% ROI)

**Most programs have not achieved the full potential ROI** due to value leakage

**New potential: Implementation of generative AI cloud programs to reduce value leakage,** net ROI impact, %



**30–40% increase in cloud foundations** (establishing foundational LLM models / adopting generative AI solutions, new recurring cloud platform spend[1])

**30–40% lower cost of migration and modernization of workloads for cloud,** driven by generative AI–enabled discovery, spec engineering, code refactoring, QA/test

**20–30% timeline acceleration in migrating/ modernizing workloads** to cloud, mitigating stalled migration value leakage drivers

**30–50% faster new application development/** feature release, driven by tools such as GitHub Copilot, StarCoder

**Increase in overall cloud program ROI,** driven by generative AI adoption across activities

[1] Includes new software licensing for generative AI platforms (such as Copilot) and additional infrastructure cost in compute/storage.
[2] There is potentially even more ROI gain through new business use cases unlocked by generative AI on cloud (eg, sales, go-to-market, customer user experience/care), driven by emerging smart agents and LLMs.
Source: CloudSights

Part III

# What actions should we take?

The ten essentials

**Having a broad understanding** of cloud's value potential and how ROI economics work is crucial for building commitment and conviction. But actually attaining that value and ROI requires action. In this section, we highlight ten actions that are essential for any cloud program:

## Discovering the full value of cloud

1. Ground the cloud strategy and program design in business value
2. Optimize the on-premises environment
3. Migrate complete customer journeys or business domains

## Solving critical technical problems

4. Build cloud foundational services to support developer innovation, scale, and security
5. Employ security as code to reduce misconfiguration without reducing agility
6. Manage data for improved cloud performance and lower costs
7. Migrate mainframe workloads when there is a business need

## Delivering organizational change

8. Adopt an agile product and platform operating model
9. Modernize applications to get the full benefit of cloud services
10. Use FinOps to control and optimize cloud spend

# Discover

## 1 Ground the cloud strategy and program design in business value

Companies that have captured the most ROI from cloud consistently focus on high-value use cases aligned to their overall business strategy. They have a clear view of cloud economics and a deep understanding of how cloud can deliver the greatest value to the business beyond simply reducing IT costs.

**Why this is important**

The first and perhaps most important action that organizations should take is to build a cloud investment strategy in the area that promises the most potential value: business enablement. On average across industries, the business value of cloud is two and half times larger than its IT value.

Beyond providing a much larger portion of value, business enablement is the area where organizations are best able to establish a distinct competitive advantage over their peers. The IT cost savings from cloud are real, but business-enabling cloud use cases can build proprietary value by expanding an organization's capabilities, speeding the pace of innovation, and driving new areas of growth.

As discussed earlier, a properly designed cloud program can drive more than 180 percent ROI over about six to eight years. In contrast, organizations that don't design a program strategically can quickly see erosion in value, with some programs even losing money. Furthermore, choosing a fit-for-purpose cloud service provider, and a well-executed sourcing process can lead to a difference of 15 to 30 percent or more in cost savings for cloud hosting in the form of service discounts, training, and credits for professional services, as well as additional strategic partnership benefits such as early access to features or flexible payment terms.

**State of the industry**

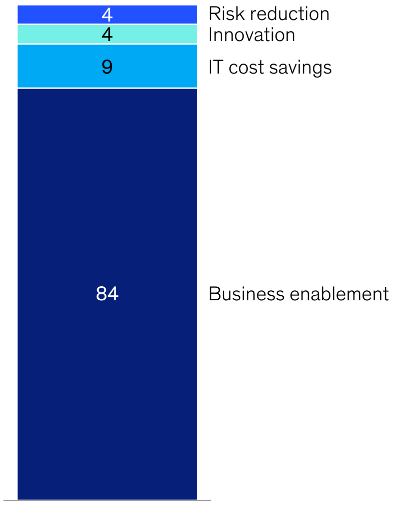
Roughly 90 percent of the executives we interviewed cited business enablement as their primary driver for moving to cloud (Exhibit 16). Organizations with mature capabilities find the speed and agility of cloud allows faster deployment of new capabilities and ingestion of new data sources.

A first step for many companies, regardless of industry, is to use cloud to eliminate a costly data center. Over time, as experience and capabilities grow, companies look to business-enablement use cases centered on adopting cloud through SaaS, analytics, and modernization. The more-advanced businesses focus on more-advanced use cases, such as building new cloud-enabled businesses (Exhibit 17).

Exhibit 16

**Business enablement is the primary reason for adopting cloud.**

**Primary driver of cloud adoption,** % of companies (n = 56)



Note: Figures do not sum to 100%, because of rounding.
Source: CloudSights

Exhibit 17

## Companies typically start with cloud to save on costs but shift their focus over time to value-generating opportunities suited to their sector.

**Value progression over time**

| | Cost savings | Business innovation/enablement and new business building | |
|---|---|---|---|
| **Banking** | • Data center exit<br>• Risk management<br>• Corporate function apps<br>• SaaS apps<br>• Digital workforce apps<br>• Contact center apps/chatbot | • Settlement processing<br>• Mainframe modernization<br>• Customer loyalty program<br>• Fraud management<br>• Investment management tools<br>• Mobile apps<br>• Auditing system | • Climate-risk modeling<br>• Mortgage origination<br>• Open bank API<br>• Wealth-management platform<br>• Omnichannel banking platform<br>• Marketing automation<br>• Customer-engagement platform |
| **Life sciences** | • Data center exit<br>• First-level support<br>• Field-force scheduling<br>• Product prediction demand | • Tracking sales-force performance<br>• Patient onboarding platform<br>• Single-instance ERP apps<br>• High-performance computing for R&D<br>• Predictive insights on patients | • R&D apps<br>• Identifying drug candidates<br>• Clinical research<br>• Predictive analytics in manufacturing<br>• Edge analytics |
| **Consumer** | • Data center exit<br>• HR-employee churn<br>• Route-optimization algorithm<br>• Pricing optimization | • Data/analytics platform<br>• Contact-center apps/chatbot<br>• Customer-engagement apps<br>• Loyalty management<br>• Employee scheduling<br>• Warehouse management<br>• Supply chain management | • Design to value<br>• Geospatial analytics for new store location<br>• New e-commerce platform<br>• Advanced inventory modeling<br>• Smart manufacturing |
| **Insurance** | • Data center exit<br>• Risk management<br>• Lapse-prediction model<br>• Data modeling<br>• Data storage and analytics | • Contact-center apps<br>• Next product to sell<br>• Partner integrations<br>• Customer portal<br>• Automated underwriting<br>• Financial advisor product | • Greenfield benefits business<br>• Fraud management<br>• Claims management<br>• Policy and benefits administration |
| **Advanced industries** | • Data center exit | • Enterprise apps<br>• Demand forecasting<br>• High-performance computing for apps<br>• Analytics for predictive maintenance | • Product engineering<br>• Mobility apps<br>• New e-commerce platform<br>• Autonomous driving systems<br>• IoT/edge analytics<br>• Supply chain management |

Source: CloudSights

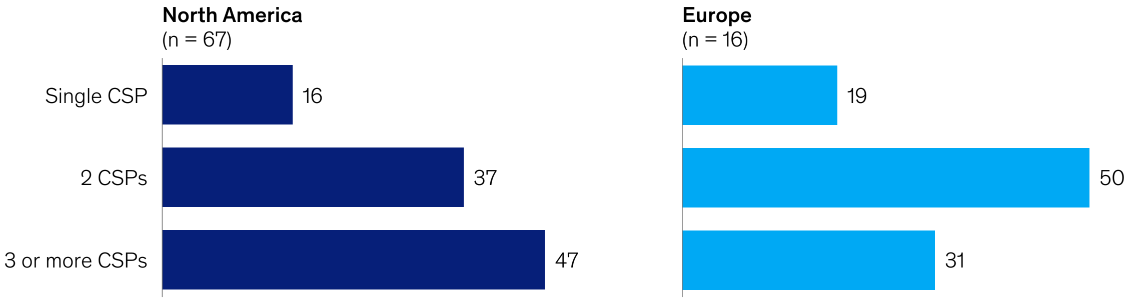
More than 80 percent of organizations in Europe and North America use two or more cloud providers, reflecting the fact that large organizations have a heterogeneous mixture of applications, infrastructure, talent, and skill sets that predispose them to one CSP over another (Exhibit 18).

Organizations that choose a single CSP are often seeking simplified maintenance, faster talent upskilling, specific tools and platform capabilities, and lower costs (through better scale discounts and lower network egress and operational costs). Organizations often choose a second CSP to take advantage of its particular strengths or to increase their own competitive leverage. But they must be ready to invest in upskilling talent to the second CSP's system and in maintaining architectural and operational consistency to allow for easy transfer of workloads.

Exhibit 18

## Organizations have different reasons to adopt multiple CSPs but often choose one as primary.

**CSP adoption,** % of companies



Source: CloudSights

Given this multicloud reality, the more important question is how organizations choose to distribute their workloads across CSPs. Our surveys indicate that organizations often choose a primary and secondary CSP, migrating core workloads to the primary provider (Exhibit 19).

### Components of success

— **Partner with the business.** Since about 75 percent of cloud's benefit is found in generating and accelerating value for the business, it's the business that needs to lead the change in cloud. This means having accountability, taking ownership, and providing funding. In practice, it requires a mechanism to support frequent interactions between IT managers and their counterparts in business units, especially those who own products and capability areas. Business leaders need to appoint knowledgeable decision makers as owners or co-owners for each priority domain or product.

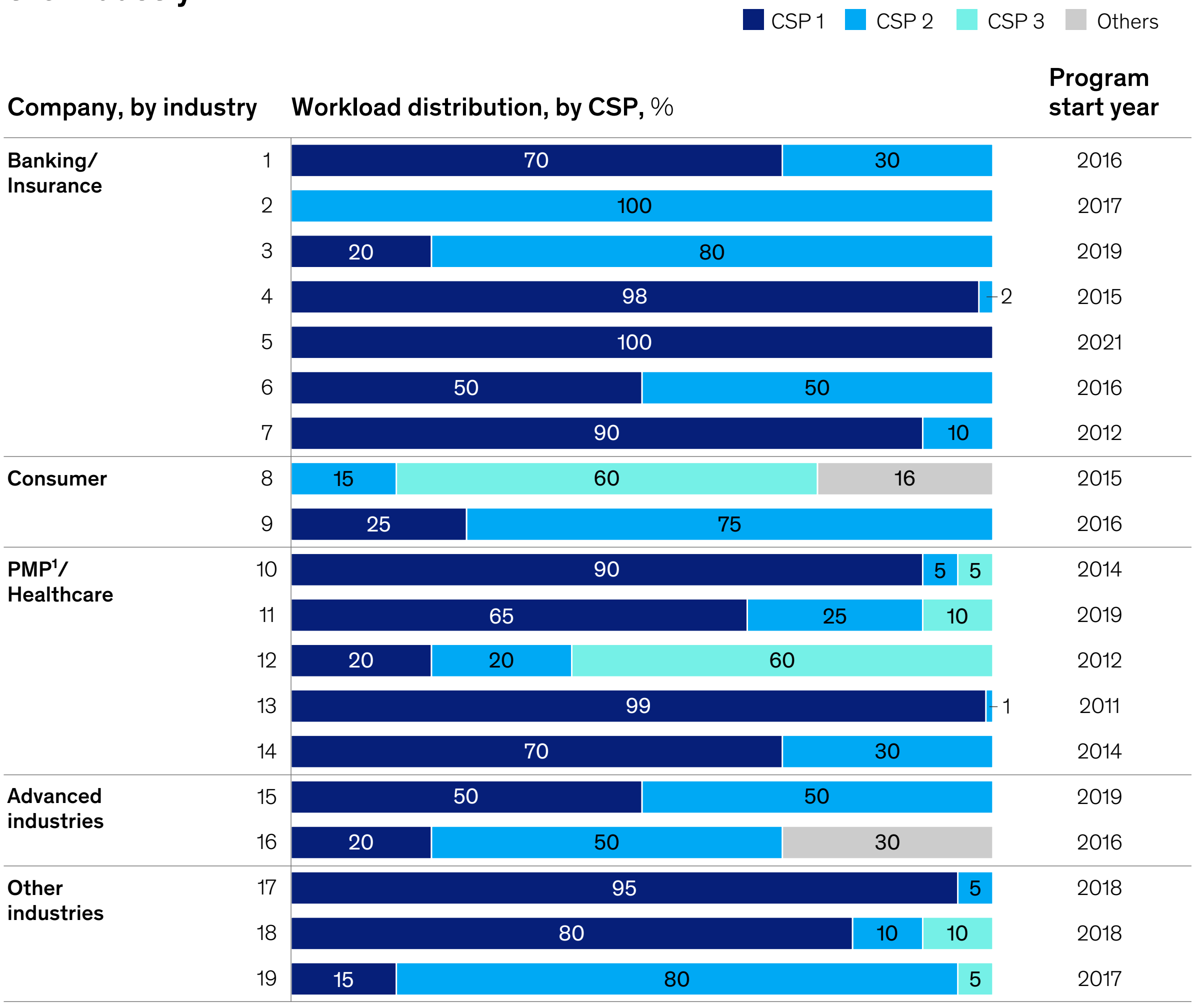
— **Set a clear vision and specific objectives.** Clearly articulating the business vision and objectives that cloud will enable helps establish the right priorities, investment levels, and focus for business, app development, and infrastructure teams—which is critical given the long-term investments and time horizons typical of cloud transformations.

— **Develop a business case and prioritize use cases.** Invest in a robust business-case development process with a clear view of the use cases to be enabled. When well executed, the business case is based on a clear understanding of both the unit economics of cloud and how the initiatives support the business's broader digital and AI transformation program, for which cloud is a fundamental enabler. Without that level of tight connectivity and coordination, even promising business cases won't deliver on the business's priorities.

— **Design architecture to scale cloud platform services.** Once the business and technical objectives have been established, organizations should make key architectural design decisions up front and invest in building flexible and scalable cloud foundational platform services. A methodical approach to technical design will help organizations avoid not only cloud sprawl, which can reduce ROI by 65 to 75 percent, but also costly rebuilds and duplication.

Exhibit 19

## Cloud workload distribution across CSPs varies significantly, depending on the industry.



[1] Pharmaceuticals and medical products.
Source: CloudSights

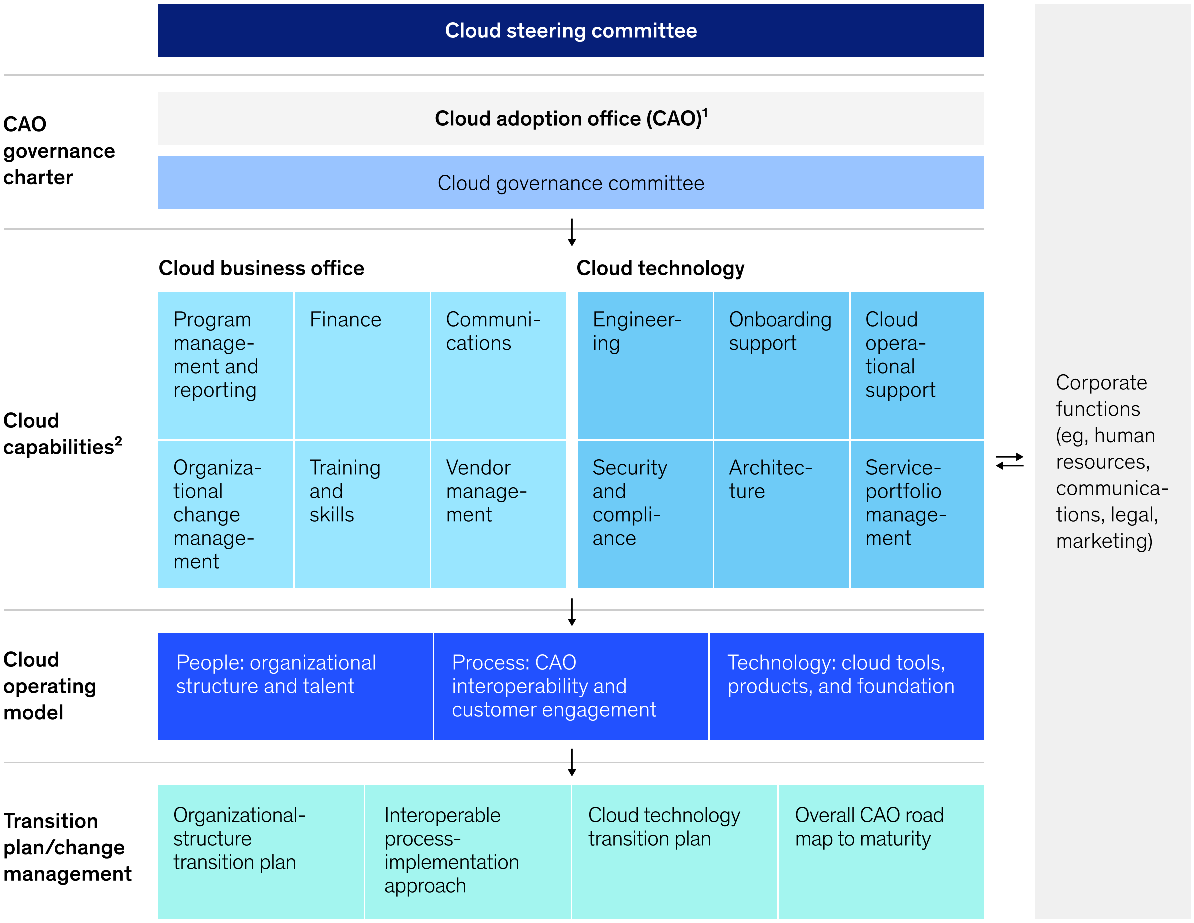
— **Establish a cloud adoption office.** Executing and sustaining a cloud transformation requires changing an organization's technology *and* operating model. Cloud transformations require well-coordinated action across all application teams, infrastructure teams, and businesses impacted by migrations. A cloud adoption office helps provide an operating model for communication and collaboration across the entire cloud program (Exhibit 20). As we will describe later, part of this change also means shifting to an agile or product-based operating model to take advantage of cloud's speed and agility.

— **Evaluate and choose CSPs.** A well-defined CSP sourcing strategy starts with understanding the business's needs and the CSP's technical capabilities and services. Typically, assessing CSP contract pricing requires careful consideration of future spend at subservice levels to unlock discounts not offered at cross-service levels. To maximize value, organizations must also factor in specific advantages and needs beyond costs and discounts, such as talent augmentation, sandbox environment accounts, and marketplace benefits.

Exhibit 20

## A cloud adoption office provides an operating model for communication and collaboration across the cloud program.

**Illustrative**



**Cloud steering committee**

**CAO governance charter**

**Cloud adoption office (CAO)**[1]

Cloud governance committee

**Cloud capabilities**[2]

**Cloud business office**

- Program management and reporting
- Finance
- Communications
- Organizational change management
- Training and skills
- Vendor management

**Cloud technology**

- Engineering
- Onboarding support
- Cloud operational support
- Security and compliance
- Architecture
- Service-portfolio management

Corporate functions (eg, human resources, communications, legal, marketing)

**Cloud operating model**

- People: organizational structure and talent
- Process: CAO interoperability and customer engagement
- Technology: cloud tools, products, and foundation

**Transition plan/change management**

- Organizational-structure transition plan
- Interoperable process-implementation approach
- Cloud technology transition plan
- Overall CAO road map to maturity

[1] We are using the term cloud adoption office (CAO) instead of center of excellence, as this better represents the office's full set of capabilities and value.
[2] As cloud maturity increases, certain capabilities may be deprecated and merge with traditional IT organization functions.

These practices can increase the likelihood that an organization's cloud strategies will lead to true business wins. A global pharmaceutical company, for example, established a cloud adoption office to help manage its $500 million-plus cloud program. It developed an automated executive dashboard that tracked value generation from budget overruns to operational progress for migration and data center shutdowns; brokered a three-way deal with a systems integrator and a CSP to ensure timely app migration and modernization and access to cutting-edge cloud-build capabilities (such as a data warehouse for enhanced security and data quality); and structured standardized terms for payments, investments, and credits with vendors. This enabled the office to quickly identify and address $30 million in overruns and more than 3,000 "orphan" servers that could be affected by migrations and take steps to remediate them.

## 2 Optimize the on-premises environment

Many organizations have ambitious public-cloud adoption goals. A significant part of their technology environment, however, will continue to run outside public cloud for the next few years. It is critical to optimize the on-premises cloud environment to provide increased agility to the business while controlling costs.

**Why this is important**

One of the most important contributing factors to the three- or four-year time frame to break even on new cloud programs is the need to maintain an on-premises environment in parallel. Organizations need to aggressively manage their on-premises technology to reduce costs, help fund the cloud transition, and strengthen the overall business case. We have seen optimization of the on-premises environment provide 20 to 30 percent productivity gains while also improving agility and resiliency.

**State of the industry**

While shifting workloads to public cloud is a popular topic, public-cloud spending is still lower than spending for on-premises IT infrastructure. According to Gartner, public-cloud spending won't overtake on-premises spending until 2025.[10] Even after 2025, on-premises spending will likely remain a significant portion of an organization's technology footprint. For this reason, we expect to see technology leaders devoting significant resources to optimizing their on-premises infrastructure.

This will require understanding the interdependencies between the two environments and determining how to best align resources and activities to maximize the value from both. Such coordination should be part of an overall technology transformation tied to business strategy. Our research shows that companies that lead in technology are better at both tying their transformation efforts to the company's overall strategy and coordinating a broad set of interrelated technology activities. We have found, in fact, that they have a greater than 60 percent likelihood of scoring high in ability to work on multiple initiatives, compared to about 25 percent for technology laggards.

**Components of success**

— **Standardize the solution offerings for on-premises infrastructure.** One of the most important features of public-cloud services is their standardization, which helps CSPs provision them rapidly. Organizations need to adopt the same concept in their solutions catalog, limiting options and customizations to improve speed and reduce costs. For example, a large technology organization realized that they provisioned about 1,000 unique combinations of developer environments during a year. They simplified their solutions catalog to fewer than 25 standard options, which improved their provisioning speed from a few days to a few hours.

— **Establish FinOps practices for the on-premises environment.** The complex cost structure of public cloud has led many organizations to establish a dedicated FinOps capability (see the last section in Part III). This approach has helped them optimize costs for both cloud and on-premises environments covering all major solutions or technologies.

Traditionally, organizations use a variety of standard options to keep costs down, such as improving the utilization of servers, reducing overallocation of storage, or optimizing storage tiers. But we find that companies can further optimize costs by 10 to 15 percent through FinOps practices that improve asset utilization or allocation.

— **Invest in newer technologies to enable efficient hybrid-cloud operations.** To address the complexities of hybrid-cloud operations, organizations have seen value in technologies that enable hybrid-cloud networking, such as SD-WAN and SASE, and in data fabrics that enable sharing of data across hybrid environments. These technologies

[10] "Gartner says more than half of enterprise IT spending in key market segments will shift to the cloud by 2025," Gartner press release, February 9, 2022.

can be considered on a case-by-case basis with a clear view of the trade-offs between the value and expense of hybrid-cloud interoperability.

- **Automate production support.** Organizations spend 30 to 40 percent of their overall IT budget on production management across service desk, help desk, monitoring, application maintenance, and other IT service management (ITSM) processes. While many organizations have implemented lean-IT principles, AIOps, and other automation technologies to streamline processes and reduce costs, many processes and operations remain manual.

  For example, a large life insurer identified a range of automation opportunities—ticket routing, recycling of servers, auto-patching, predictive monitoring, classification of incidents, and higher first-contact resolution (FCR) for simpler incidents—across its 2,200 process steps that can reduce its production support budget by 15 to 20 percent.

## 3 Migrate complete customer journeys or business domains

Capturing value from one activity often requires changing related activities as well. Improving the process for opening a bank account, for example, won't yield much value if the process for managing that process isn't also improved. It's therefore critical to think in terms of complete customer journeys or domains to account for all the necessary dependencies. This approach helps companies identify and migrate clusters of related workloads to ensure that the entire customer journey gets value from cloud.

**Why this is important**

Many leaders, especially those under intense pressure to justify their cloud investments, will be tempted to focus on short-term gains. They may, for example, migrate applications simply to reach a certain migration target. This short-term focus can frequently lead to bigger headaches later, such as greater difficulties and limited value in migrating other applications. The problem is that migrating a random set of unrelated applications chokes off the broader value opportunity.

A better approach is to prioritize the migration of a set of applications needed to deliver a given customer journey. This doesn't mean that every application or element of the customer journey needs to be migrated to cloud. But it does ensure that important dependencies are addressed so that migrating those apps generates business value.

When reviewing which applications to migrate, a company shouldn't end with a yes or no decision. Leading companies take it as an opportunity to improve the overall customer journey or the process itself. One insurer, for example, improved its claims process by enabling customers to submit pictures they took of an accident to the company's app. The images were uploaded to cloud, where AI-driven analysis assessed damages and repair costs. Customers were offered a settlement within three minutes, which they could accept or dispute. Some 60 percent accepted, which saved the company $1 billion in claims-related costs in the first year.

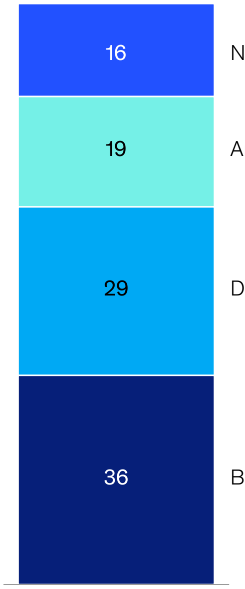
**State of the industry**

Organizations today weigh a range of factors when prioritizing applications to migrate (Exhibit 21). Many use a combination of business value and architectural readiness. However, most are still analyzing applications individually rather than systemically, based on business domains or customer journeys.

Exhibit 21

## ‘Business value’ has become companies’ most important consideration in migrating apps.

**Considerations when migrating apps,**[1] % of companies (n = 79)



[1] Q: *How do you prioritize applications for migration?*
Source: CloudSights

### Components of success

- **Identify applications related to priority journeys.** The business needs to play an active role in setting the migration program, starting with understanding the value of individual applications and how they serve the corresponding customer journey. This clarity helps to identify and prioritize which application sets to be migrated. Asking which business domains (such as order capture, billing, or supply chain optimization) would benefit most from the speed, innovation, and scalability that cloud platforms can provide can help business leaders identify the highest-priority areas for migration to cloud. The outcome should be a prioritized, sometimes multiyear road map of domains in which cloud will accelerate performance and digital transformation.

- **Determine needed components and dependencies.** It is often difficult to know the dependencies within applications and systems, even for companies with advanced capabilities. There are now tools available to automate and accelerate the discovery of application dependencies throughout a portfolio beyond what is possible using a company’s configuration-management database (CMDB). In one organization, we found that over 80 percent of the applications across customer journeys had other applications they were dependent on or that were dependent on them. Identifying these dependencies helps companies understand which applications to “bundle” for migration.

- **Evaluate applications by value and migration complexity.** Not every application should necessarily be migrated to cloud. Companies need to evaluate each one by the associated business case and complexity to determine its best disposition—rewrite, replatform, retire, leave on-premises, or migrate. This exercise also identifies actions needed to enable supporting capabilities such as platforms, continuous integration and continuous delivery (CI/CD) processes, services, and so on.

- **Group applications and plan waves of migration.** Use the understanding of dependencies and dispositions to determine which groups of apps must move together in waves of migrations and when. This approach provides flexibility to adjust if, for example, business priorities change while applications are awaiting migration, and significantly increases the chance of migration success. In our earlier banking example, grouping dependent applications to migrate together eliminated latency and data concerns.

# Solve

## 4 Build cloud foundational services to support developer innovation, scale, and security

Cloud architecture needs to scale up as cloud workloads grow, and it needs to empower developers to build resiliency and security and meet compliance requirements. The right cloud foundation provides ready-to-use, precertified, configurable solutions that result in faster and more cost-effective transformations.

### Why this is important

At an organizational level, a strong cloud foundation can shorten product development and deployment times. For example, a large payment processor was able to introduce an AI-powered risk engine in just 13 weeks "from idea to live." The key was having cloud products, including AI services, available as part of the cloud foundation to give development teams a head start on their build efforts.

A well-designed cloud architecture needs to meet the needs of three important IT practitioner groups:

- **Developers**, by providing a set of cloud products that offer the infrastructure, deployment pipelines, code repositories, and security controls necessary to begin building cloud applications

- **Site-reliability engineers (SREs)**, by automating standard migration tasks and reusing precertified components to accelerate tasks and reduce app migration costs. Once apps are in cloud, foundational services reduce the time needed to maintain and patch them so that SREs can support more environments than they could on-premises

- **IT infrastructure and security teams**, by enabling a federated approach to ensuring resiliency and security that defines patterns and builds automation centrally for use by application development teams as needed

Often organizations will begin by focusing on a limited set of applications to be migrated or modernized in cloud. This often achieves short-term goals but doesn't provide any capabilities that carry over to the next round of migration or development. Focusing first on the foundation needed to support the overall transformation benefits all business domains.

### State of the industry

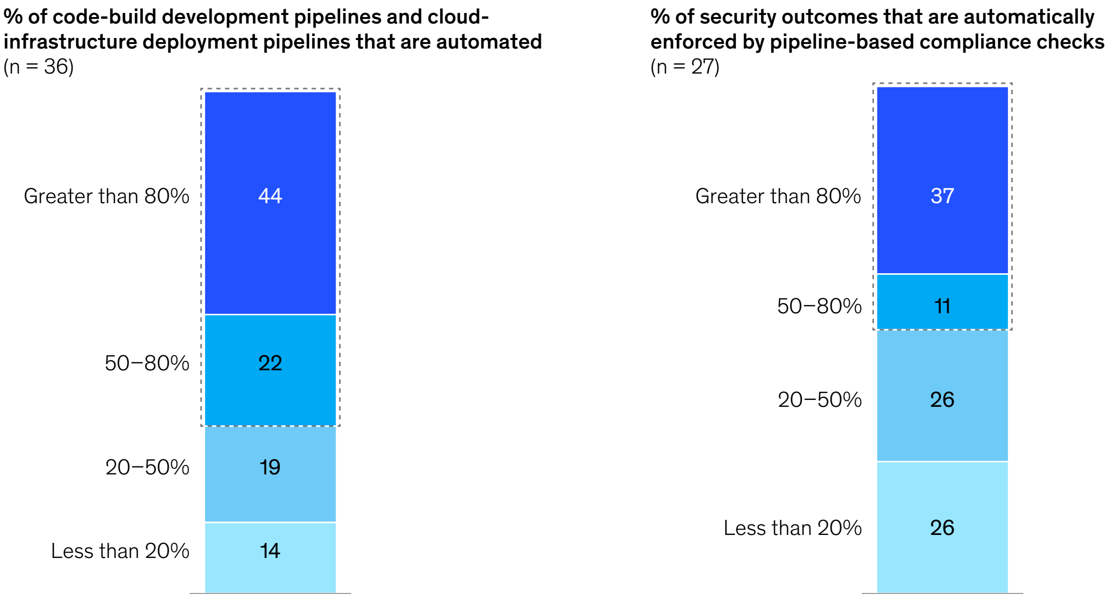
Half of the companies we've surveyed are moving toward infrastructure as code, though automated full-service deployment is generally still not a reality. Companies in general are further along on their infrastructure automation than they are on security automation (Exhibit 22).

In terms of resiliency (Exhibit 23), we observe that more than 75 percent of organizations have adopted a multiregion, multizone approach and have enabled active-active replication for tier-one applications.

Exhibit 22

## Almost two-thirds of organizations surveyed have automated more than half of their development and infrastructure pipelines.

**Level of automation,** % of companies

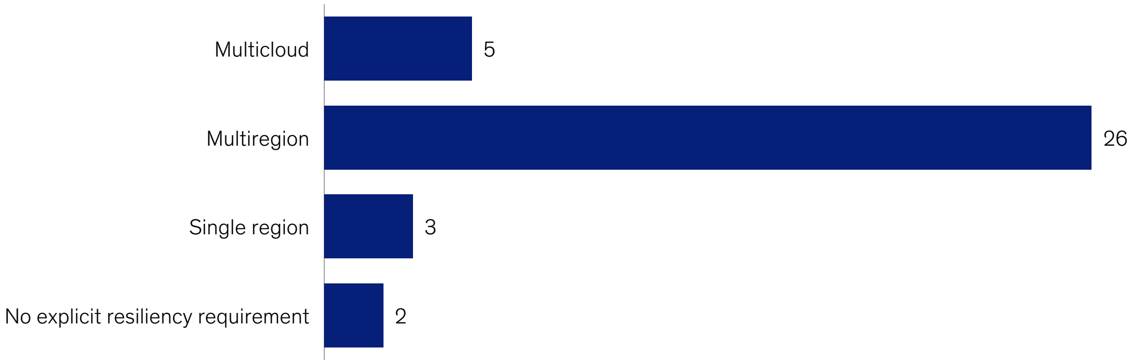


Note: Figures may not sum to 100%, because of rounding.
Source: CloudSights

Exhibit 23

## A multiregion architecture continues to be the preferred resiliency approach for tier-one apps.

**Preferred architecture resiliency approach,**[1] number of companies



[1] Q: *How are you architecting resiliency for tier-one apps when migrating to and scaling in cloud?* (n = 36).
Source: CloudSights

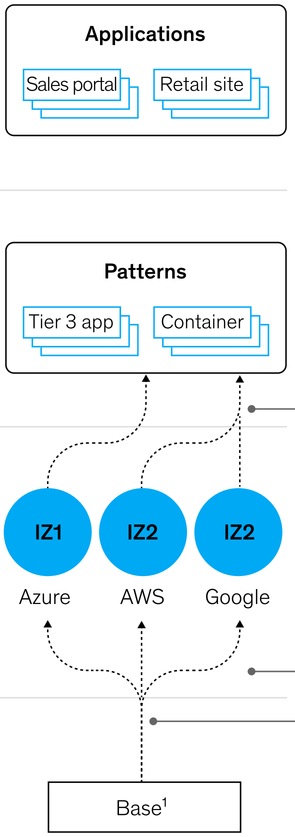
### Components of success

Establishing a robust cloud foundation consists of defining and implementing three main layers with a single foundation that can handle multiple CSPs. This point is critical and worth emphasis. You want to put in place and configure a foundation that allows you to take advantage of the services and capabilities each CSP offers (Exhibit 24).

Exhibit 24

## A cloud foundation architecture allows for tailored security controls, easy scalability, and appropriate resiliency.

| | |
|---|---|
| **Applications**<br>Organizations purchase and build applications to enable business outcomes; business product owners focus on how the app benefits the organization | → **… prioritize business outcomes** by separating the logic of the application from the infrastructure that hosts it |
| **Application patterns**<br>Used by application developers to deploy in a consistent way; inherits the governance defined in lower layers to ensure compliance | → **… reduce up-front developer work while providing flexibility** through standardized patterns that can be adapted to specific use cases |
| **Isolation zones (IZ)**<br>Logical separation per modernization level and/or specific requirements (eg, data classification); inherits policy of base below it, and adds more granular IZ-specific policies | → **… enable secure migration of unmodernized apps** by segmenting them into subdivisions that customize controls rather than the one-size-fits-all, cloud-native approach today |
| **Base**<br>Policies that apply to all isolation zones and application patterns; foundational services (networking, identity, logging and analytics, etc) | → **… accelerates cloud migration by automating foundational services,** which prevents duplication of work across the enterprise |
| **Foundation** | |



Through pattern configuration, applications can be quickly deployed in the cloud

Automatic inheritance of common controls reduces redundant configuration for every account

More-consistent control enforcement due to fewer control points

[1] A single primary base services most isolation zones, but additional bases (eg, an integration base for M&A activities) can also exist in the architecture.

— **Build out the three layers of the cloud foundation**

1. **Application patterns**, whose reusable components enable consistent and streamlined app development, are code artifacts that automate secure, compliant, and standardized configuration and deployment of apps with similar functional and nonfunctional requirements using:

   » **Policy as code:** the translation of an organization's standards and policies into executable code that automatically secures the infrastructure and environment of the organization in accordance with policy

   » **Security as code:** software that verifies the configuration of an infrastructure's actual definition before and after deployment to meet a particular defined standard (see more on this later)

   » **Compliance as code:** a composed set of rules interpreted by a software-based policy engine that enforces compliance policy for a specific cloud environment

   » **Infrastructure as code:** practice by which infrastructure teams use software-development engineering practices, work methods, and code to provision environments and move beyond the inflexible "hardware" mindset

2. **Isolation zones** are a set of separate CSP-specific zones (sometimes called landing zones) that isolate application environments to prevent risks from spreading. Each zone contains CSP services, identity and access management (IAM), network isolation, capacity management, shared services scoped to that isolation zone, and change controls where one or more related applications run. These zones are helpful in that they deliver CSP-specific capabilities and isolate applications and data to reduce the spread of security breaches and failures.

3. **The base** is a set of CSP-portable capabilities provided to a set of isolation zones, including network connectivity and routing; centralized firewall and proxy capabilities; identity standardization; enterprise logging, monitoring, and analytics (ELMA); shared enterprise services; golden-image (or primary-image) pipelines; and compliance enforcement.[11]

A key design element is CSP-portable solutions (such as Terraform) for base and isolation zone layers, to enable greater flexibility.

— **Integrate foundational services for generative AI into the overall cloud architecture**

The complexity and risks associated with generative AI require developers to expand their range of foundational services. They will need to use a common set of foundational capabilities, including an experience layer, model hubs, foundational models, DevOps tooling, and data platforms. See Exhibit 25 for the full stack of capabilities needed to enable generative AI on cloud.

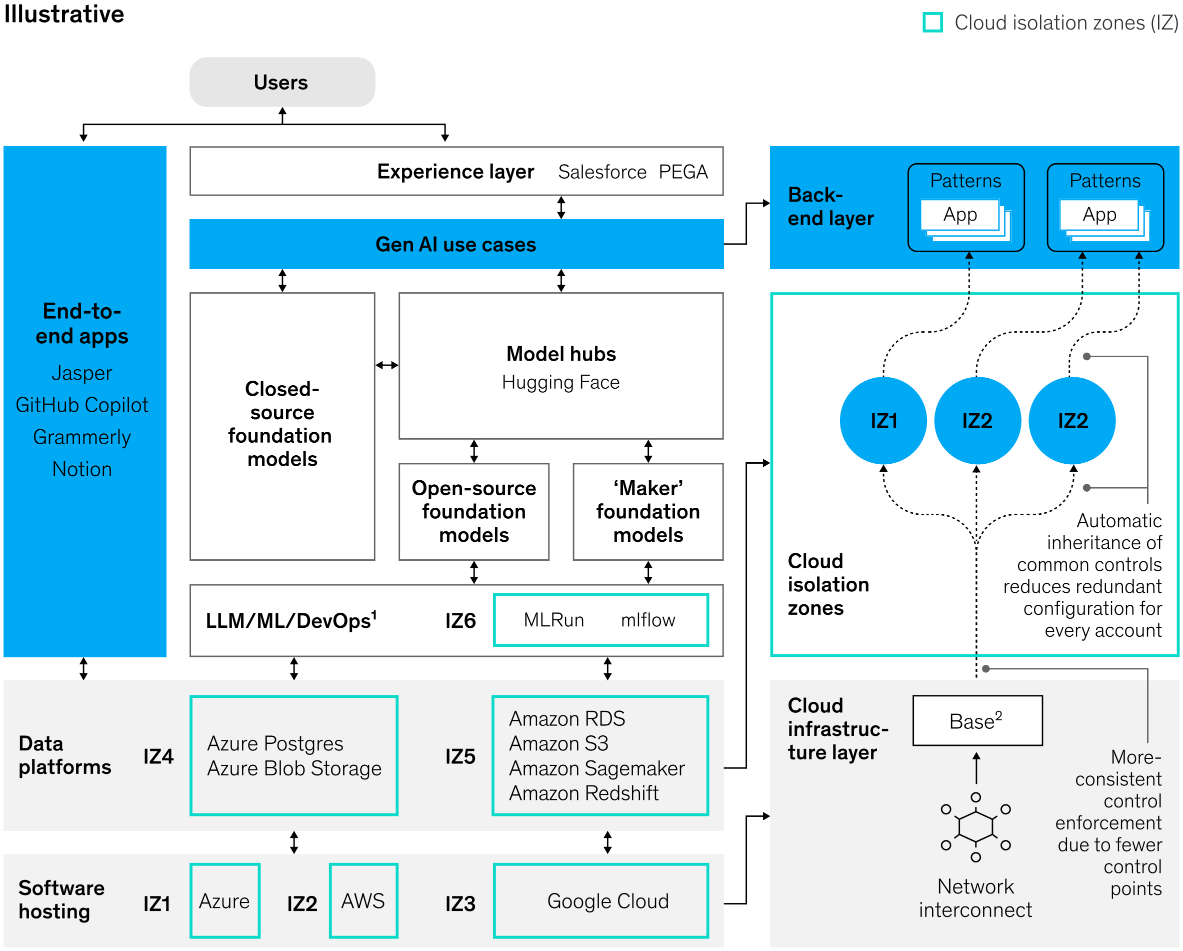
In developing the cloud foundation, organizations need to integrate resiliency capabilities into each component. The base, for example, needs to be resilient itself but also provide resiliency features for the isolation zones that use it. This requires a clear business-first understanding of the various resiliency needs and priorities and how they can be embedded and invoked in the right layers of the foundation and in the applications themselves. Making that determination requires careful consideration from leaders, who must do three things:

[11] Descriptions of the three layers can be found in Aaron Bawcom, Sebastian Becerra, Beau Bennett, and Bill Gregg, "Cloud foundations: Ten commandments for faster—and more profitable—cloud migrations," McKinsey, April 21, 2022.

Exhibit 25

## The right cloud foundation for generative AI involves an architecture connecting the back end, data, and cloud infrastructure.

**Illustrative**



[1] LLM = large language model; ML = machine learning.
[2] A single primary base services most isolation zones, but additional bases (eg, an integration base for M&A activities) can also exist in the architecture.

1. **Understand business resiliency requirements for key journeys.** Technology and business leaders should jointly identify the most-critical business journeys and their associated applications. Applications can be categorized into four levels: mission critical, business critical, business operations, and administrative (Exhibit 26).

2. **Assess critical journeys and map resiliency patterns.** Start by identifying existing vulnerabilities and technical constraints in the systems for each journey. Assess how they affect the customer experience and what fixes are needed in infrastructure, application architecture, data management, or monitoring and tooling. Based on this, map a set of relevant resiliency patterns to address gaps. Exhibit 27 summarizes the 13 most-relevant resiliency patterns (seven that can be implemented based on application criticality and six that can be implemented for all applications).

3. **Prioritize patterns and develop an implementation road map.** Prioritize resiliency patterns based on required investment, implementation complexity, timeline, and tradeoffs. The output should be a road map for implementation.

Exhibit 26

## Prioritize resiliency by application or journey criticality.

### Application/journey criticality levels

| | Level 1:<br>Mission critical | Level 2:<br>Business critical | Level 3:<br>Business operational | Level 4:<br>Administrative |
|---|---|---|---|---|
| **Description** | Requires continuous availability; breaks in service are intolerable immediately and significantly damaging; availability required at almost any price | Requires continuous availability, but short outages are not catastrophic; availability required for effective business operation | Contributes to efficient business operation but not in direct line of service to customer | Office productivity tools for business to operate; failures do not affect customers |
| **Typical characteristics** | **RTO[1] or RPO[2] near 0 minutes**<br>• generates revenue<br>• external customers are direct users<br>• underpins (eg, is shared platform for) several other journeys | **RTO[1] < 30 minutes, RPO[2] 0–15 minutes**<br>• indirectly affects revenue generation<br>• supports activities essential for effective business operation<br>• organization-wide dependency/ pervasiveness | **RTO[1] 30–60 minutes, RPO[2] 2–4 hours**<br>• supports operational activities<br>• mostly internal users<br>• affects efficiency and cost of operation | **RTO[1] 1–4 hours, RPO[2] 4–8 hours**<br>• used exclusively internally<br>• supports individual productivity |
| **Unavailability impact** | Direct impact on public/ national safety<br>Immediate damage to reputation<br>Damage to revenue generation<br>Regulatory penalties and noncompliance | Indirect impact on public safety<br>Prevents collection of revenue<br>Significant impact on customer service<br>Significant disruption to operation | Reduced efficiency and increased cost of operation | Reduced individual performance and productivity |
| **Criticality** | High | | | Low |

[1] Recovery time objective.
[2] Recovery point objective.

For example, one North American payment processor followed these steps to reduce technical debt and address stability issues while replatforming its core API payment gateway. As a result, it was able to implement an active-active solution across two regions; a canary deployment approach (employing a Canary API); asynchronous replication with eventual consistency; a mixture of serverless and autoscaling servers; and redirection of all transactions to a new payment gateway. These enhancements made deployment four times faster, led to near-zero downtime, and lowered the cost of operations.

Exhibit 27

## These 13 patterns provide cloud resiliency.

**Infrastructure resiliency patterns based on application criticality**

Cost or complexity of implementation
High ▨ Low

| Dimension | Resiliency patterns | Tier 1: Mission critical | Tier 2: Business critical | Tier 3: Business operational | Tier 4: Admin-istrative |
|---|---|---|---|---|---|
| Architecture design | 1 Multicloud[1] | Multicloud or single cloud | Single cloud | | |
| | 2 Multiregion/ multi-AZ[2] | Multicloud or single cloud | Warm standby | Pilot light | Back up and restore |
| | 3 Elasticity | Serverless or autoscaling servers | | | |
| | 4 Load leveling | Competing consumer or queue based | | Queue-based | Request throttling |
| Data management | 10 Durable data storage[3] | Synchronous or asynchronous[4] | Asynchronous | | |
| Deployment strategy | 12 Deployment pattern[5] | Blue-green or canary | | Canary | Rolling |
| Monitoring and tooling | 13 Health endpoint monitoring | Service level indicator (SLI) monitoring | | Functional monitoring | Uptime monitoring |

**Application resiliency patterns**

5 Bulkhead design

6 Circuit-breaker design

7 Scheduler agent supervisor

8 Retry

9 Compensating transaction

11 Offload read operations

No application resiliency pattern exists

[1] To be used only in special scenarios where multiregion does not provide enough resiliency; most organizations do not implement this approach.
[2] Availability zone.
[3] Varies primarily based on the use case rather than on the tier of journey/application.
[4] Asynchronous replication should be used when latency is higher and need for data consistency is low to medium.
[5] Reflects minimum required level of deployment rather than suggested level; best practice is for DevOps team to be consistent with deployment patterns across

## 5 Employ security as code to reduce misconfiguration without reducing agility

Cloud can be significantly more secure than on-premises technology—but only if appropriate security automation is in place. The easiest and best way to keep cloud safe is by reducing human error. Indeed, the vast majority of cloud security failures stem not from attacks but from misconfiguration of applications and systems.

Security as code (SaC) is the best and most efficient means of securing cloud workloads with speed and agility. SaC implements cybersecurity policies, standards, and compliance automatically through code, which is then enforced in the configuration scripts used to provision cloud systems.[12]

### Why this is important

The price of not implementing the right cloud security capabilities and operating model is often a stalled cloud program, unmanaged risks, or—at worst—vulnerabilities in cloud.

Building generative AI capabilities in cloud only expands these risks:

— **Data leaking to the public domain:** Enterprise use of generative AI may result in access and processing of sensitive information, intellectual property, source code, trade secrets, and other data through direct user input or the API, including customer, private, and confidential information.

[12] "Security as code: The best (and maybe only) path to securing cloud applications and systems," McKinsey, July 22, 2021.

- **Plausibility risk:** Actors may use models or cause models to be used in ways that will expose confidential information about the model or cause the model to take actions that are against its design objectives.

- **Non-secure code generation:** Code generated by generative AI could potentially be used and deployed without a proper security audit or code review to find vulnerable or malicious components.

### State of the industry

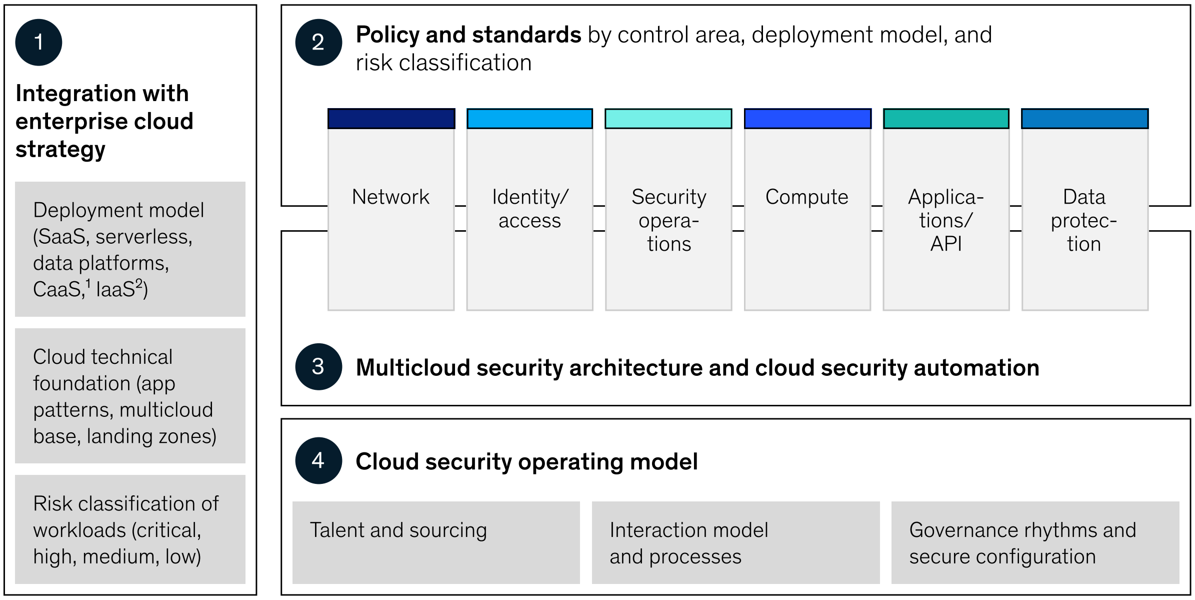
Organizations that we've surveyed indicate that they are in the early stages of using SaC. Fewer than 30 percent of organizations have implemented automated pipelines that enforce more than 50 percent of the outcomes tied to their security policies.

To bring risk and security into their cloud operating model, many organizations have deployed tooling from a growing number of vendors offering security management and policy as code. However, most are still struggling to integrate this tooling into their enterprise cloud strategy, establish the right policies and standards, build multicloud security architecture and automation, and change their security operating model to enable the desired benefits. Organizations need a comprehensive approach to establish holistic cloud security, which means different parts of the business working closely together (Exhibit 28).

Exhibit 28

## An effective cloud security framework requires four elements.

**Cloud security levers**



1 **Integration with enterprise cloud strategy**

Deployment model (SaaS, serverless, data platforms, CaaS,[1] IaaS[2])

Cloud technical foundation (app patterns, multicloud base, landing zones)

Risk classification of workloads (critical, high, medium, low)

2 **Policy and standards** by control area, deployment model, and risk classification

Network

Identity/ access

Security operations

Compute

Applications/ API

Data protection

3 **Multicloud security architecture and cloud security automation**

4 **Cloud security operating model**

Talent and sourcing

Interaction model and processes

Governance rhythms and secure configuration

[1] Containers as a service.
[2] Infrastructure as a service.

### Components of success

- **Integrate security with enterprise cloud strategy.** The design of new security controls should focus on the technologies and deployment models that are prioritized for migration. Organizations should incorporate levels of application risk into the cloud-migration schedule, starting with lower-risk apps to allow time to build in additional security controls as the cloud platform matures. This enables cloud platform teams to build isolation zones specific to architectural characteristics and security requirements, allowing like-for-like applications to adhere to the same security controls.

- **Integrate generative AI risks into the taxonomy and control model.** Risk taxonomies need to include how generative AI models can lead to impaired fairness, IP infringement, compromised data privacy, diminished data quality, malicious use, decreased performance, reduced traceability, and compliance and third-party risks. Many of these will require new types of controls, such as real-time guardrails on prompts, to prevent malicious use.

- **Set policies and standards by control area, deployment model, and risk classification.** Many organizations find that they need to update existing policies to address cloud's unique security context. Cloud often requires more nuance and specificity than traditional control frameworks. Not all controls are applicable to all workloads. Low-risk workloads don't need the same level of security scrutiny as high-risk business workloads. Control frameworks should be designed to incorporate risk classifications and deployment models, and they must be written to a level of detail sufficient to enable them to be translated into code. This is a prerequisite to establishing cloud security automation and is the only way to secure cloud applications and systems at scale.

- **Use generative AI tools to expand and accelerate implementation of SaC capabilities.** Developing compliance and security policies precise enough for automated execution requires time from in-demand cloud security engineers. Early experiments indicate that large language models (LLMs) can be used to generate and correct Terraform and Chekov policies to ensure secure container configuration.

- **Extend coverage models across all CSPs.** Holistic security means having the same coverage model across all cloud providers, ideally working toward a unified dashboard ("single pane of glass") for risk and security teams to use in measuring and managing control efficacy. Additionally, automated controls should have redundant coverage and be available at multiple points in the software development life cycle (SDLC), with checks during development, at deployment, and at runtime. Full coverage also requires integration with all CI/CD pipelines. This approach shifts more work onto security teams in the short term, even to the point of slowing them down. But the long-term payoff will be a faster, more effective organization with fewer layers of the security operating model organized around escalation, remediation assignment, and management of a continuous pile-up of known cloud security violations.

- **Establish a cloud security operating model.** An updated operating model should bring together stakeholders with clear roles and responsibilities across the SDLC. This group determines what policies are enforced as code in the deployment pipeline, identifies the root causes of common issues, reviews highly critical and open items to clear remediation blockers, tracks and reports on compliance to keep the cloud program within the enterprise's risk tolerance, and provides overall strategic oversight and alignment with the cloud program.

  Under this operating model, development teams are empowered to manage their own infrastructure. They can embrace a full DevSecOps model in which applications and infrastructure are created in tandem, and security is necessarily part of that development cycle. A next-generation SaC operating model provides developer platform integrations with clear documentation on how to make code compliant.

As an example, a US regional bank's on-premises security model didn't fit the cloud environment. Entrenched security tooling, manual review and approval processes, and limited automation increased risks. With the ongoing evolution in cloud technology in mind, the bank evaluated cloud tooling with capabilities in development of functional and declarative policies and in automated validation of policy syntax.

After choosing the tools, the bank made an enforcement plan. Automated pre-deployment checks would prevent insecure cloud infrastructure code from being deployed. This meant the bank's development teams had to implement policy as code in their software development life cycles, which couldn't be circumvented without escalating for specific approval and had to be validated through audit sampling. This approach reduced misconfigurations by 90 percent and eliminated manual configurations for the affected applications, shoring up security for the bank's multiyear cloud transformation.

## 6 Manage data for improved cloud performance and lower costs

Companies need to think through their architecture, data management (especially given widespread and varying compliance and regulatory issues), and the different cost implications of a cloud environment. CSPs offer seamlessly integrated platform services, such as centralized monitoring and logging, scheduling, and orchestration. Further offerings, such as optimized compute power and storage nodes, can help tailor compute and storage capacities to specific industries and enterprises.

### Why this is important

Data is the foundation source of value in the modern enterprise. Generative AI has only increased its importance; training LLMs requires large amounts of it, and competitive advantage from those models depends on being able to access proprietary data. Using cloud to modernize data and analytics management can increase revenue 14 percent.

While most companies have some data in cloud in some form, many do not have a modern data platform and thus end up focusing on disparate use cases. If one of their cloud-based apps needs to access on-premises data, for example, troublesome latency issues can result. Furthermore, an inefficiently organized combination of cloud and on-premises systems can quickly rack up unnecessary costs as companies are forced to pay their CSPs for data egress. Before long, the advantages of cloud—particularly performance and cost—begin to disappear.

Cloud enables disruptive business models by federating data architectures to make data easily sharable across business units and within enterprise alliances. Ecosystems and corporate alliances can tap into cloud to share selected data securely among consortium members. This increased data availability can be a catalyst for novel insights and use cases. In larger corporations, cloud also facilitates, through standardization, the organization of data architecture around distinct business units while optimally supporting a federated governance model.

### State of the industry

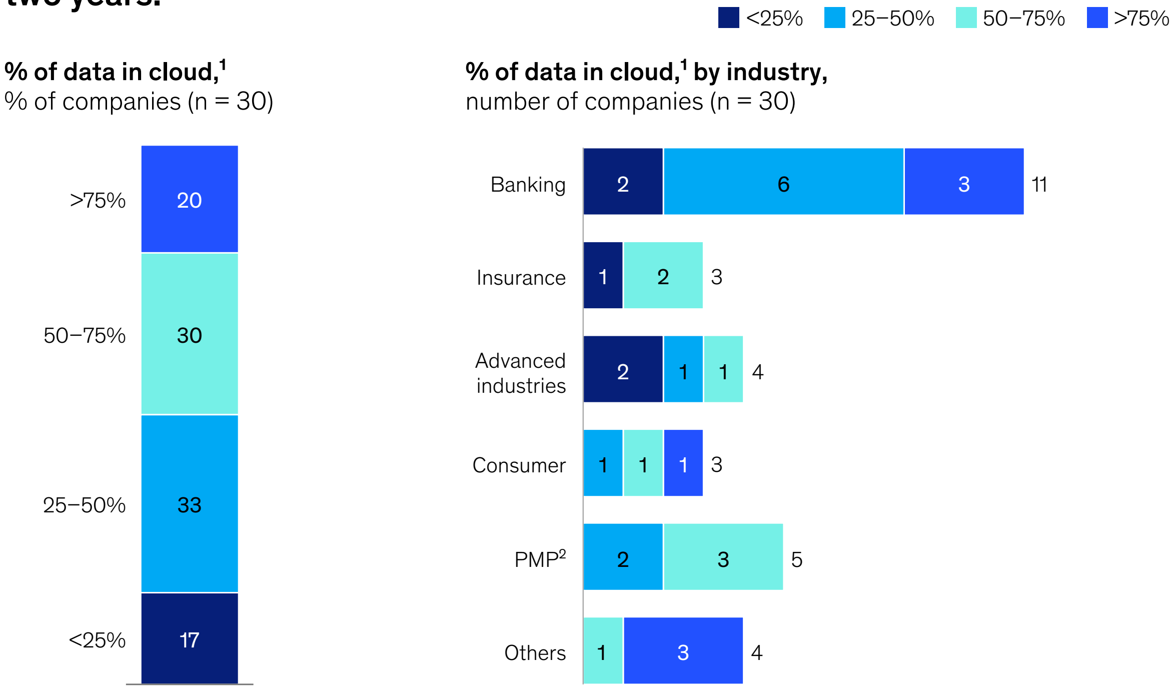
Our surveys suggest that organizations plan to move at least half of their data to the cloud in the next two years (Exhibit 29). This shift comes in the context of accelerating data-related spend, which has risen 14 percent a year from the 2017–19 period to 2020–22.[13]

Our survey further showed that most organizations don't have any specific concerns about hosting sensitive data, while "data gravity" and avoiding vendor lock-in are still issues.

[13] McKinsey Global Data Transformation Survey 2020–21.

Exhibit 29

## Most companies anticipate having at least half their data in cloud in two years.



[1] Q: *How much of your data do you see moving to the cloud versus on-premises over next two years?*
[2] Pharmaceuticals and medical products.

### Components of success

— **Group use cases into relevant domains.** Many companies focus soley on developing use cases without an overarching view of how data stores will be affected, which results in a proliferation of disconnected and uncoordinated data marts. The path to value requires grouping use cases that support a specific business goal into a domain and developing a data model that supports it. When done well, this approach can deliver value in three to six months.

— **Determine the target architecture.** Companies should do a deep dive into regulatory, data-privacy, and data-residency requirements. Assessing the feasibility of cloud migration based on their existing landscape can help companies determine the right balance of on-premises systems and cloud services. This exercise can identify the cloud-based platforms to support future data architecture and align their technology with talent availability.

— **Determine where to store and process data.** Factors such as performance, accessibility, and security should determine where to store data and how it's processed. This allows organizations to minimize the need for frequent data movement by aligning data with the appropriate infrastructure.

— **Institute agile tooling and automation.** Development teams can increase productivity by about 30 percent through DevOps, DataOps, and MLOps tooling. These offerings are available on cloud platforms and can help organizations generate value from their data. The flexibility of cloud-service provisioning enables organizations to build out a modular data architecture to serve a variety of use cases. Automation is embedded across the entire process.

- **Build for interoperability.** Ensure that systems, applications, and services can seamlessly interact with each other, regardless of the data's location. Standardized protocols, open APIs, and data formats promote interoperability and reduce challenges.

- **Employ replication and synchronization mechanisms.** This helps to distribute data across multiple locations or platforms and provides redundancy, fault tolerance, and scalability while reducing reliance on a single data center or system.

- **Establish data governance and life-cycle management.** Set policies for data retention, archiving, and deletion, as well as for ensuring compliance with local regulations such as GDPR and HIPAA, as well as security requirements to protect personally identifiable information (PII).

- **Enhance observability.** Being able to track data requires businesses to monitor it, understand where it's coming from and going to, and optimize the flow across different systems. This requires tracking the health, performance, and quality of data throughout its life cycle, from generation or ingestion to consumption and analysis.

Among the companies that have dealt with challenges around data in cloud is a major US airline. The business decided to keep its database on-premises, while moving a substantial portion (though not all) of the applications supported by the database to a public cloud. The company's intention was always to continue migrating its data and applications to cloud. In the meantime, however, the on-premises database needed to support both on-premises and cloud-based apps, which resulted in persistent latency issues. To address this problem, the company decided to establish a read-only copy of the database in cloud to create two tiers of access, thus avoiding the need for traffic to be constantly exiting cloud and helping to alleviate latency issues. The cloud-based copy is updated at periodic intervals.

## 7 Migrate mainframe workloads when there is a business need

As organizations migrate workloads and data to cloud, they must decide what to do with their mainframe. As with other cloud-related decisions, the answer should be grounded in the value at stake. In some cases, migrating parts of the software running on a mainframe to take advantage of the scalability or new capabilities in public cloud could make sense. However, in many other cases, "modernizing in place" is sufficient by, for example, tactically converting COBOL to Java code modules or elements while still running on the mainframe. This approach is especially attractive for companies whose mainframes have been heavily optimized over time.

### Why this is important

Businesses are running important systems on their mainframes, and any cloud-migration strategy needs to account for how migrated workloads can best interact with mainframe applications.

While companies are always looking for ways to modernize their mainframes, they should review the workloads in question to better understand which ones to modernize—and when—to support cloud efforts. In some cases, for example, it is less costly to modernize an app that generates data on-premises than it is to migrate the data to cloud and back. Similarly, migrating a specific module to cloud can help a workload generate more value. Using automated refactoring tools for migration and testing, for example, could help automate 30 to 40 percent of the overall effort.

### State of the industry

The process of improving the performance of mainframes is continuous. There is a spectrum of choices, from in-place application modernization on the mainframe, to migration to cloud. In the past few years, a new generation of tools and methods to improve mainframe performance has emerged. More recently, exciting generative AI tools that can help with faster and more accurate code conversion have come on the market. Code Assistant for IBM Z, for example, is designed to assist businesses in refactoring their mainframe applications.[14]

### Components of success

- **Understand the business need.** Make sure that any consideration of migrating mainframe workloads is grounded in a business need rather than just a desire to move to cloud. Often, organizations have been running core applications on mainframes for 30 to 40 years, and there are inherent risks associated with any changes to legacy core systems and data. In some cases, organizations must deal with long-term license contracts, which can be difficult to exit. Mainframe applications are also often highly dependent on one another, making it complex to migrate any single workload. Aside from technical challenges, mainframes often perform well in many use cases—and if it isn't broken, why fix it?

- **Consider a multipronged approach to migration, if needed.** If there is still a compelling case to modernize or migrate some or all of the mainframe workload(s) to public cloud, we generally recommend a multipronged approach that includes the following:

  - manual recoding of mainframe applications
  - use of automated refactoring tools that convert COBOL, Db2, and other legacy-application code and databases
  - emulation software that effectively "wraps" around the legacy code and mimics the mainframe operating system calls replication of the data to reduce dependence on the mainframe for non-mainframe systems

  With this sort of approach, organizations can transition with the proper balance of speed, cost, and reliability. Take, for instance, a large US financial services company: after launching its cloud program several years ago as part of a larger digital transformation, the company started its mainframe modernization with an intensive process of manually recoding the system. During this effort, which was projected to take years, the team realized they could accelerate the transition using automated refactoring. They are also investigating the possibility of using generative AI to aid in the process, which may eventually significantly accelerate additional modernization efforts.

[14] Kyle Wiggers, "IBM taps AI to translate COBOL code to Java," *TechCrunch*+, August 22, 2023.

# Deliver

## 8 Adopt an agile product and platform operating model

A product and platform operating model organizes technologists into two parts. One is around teams that build user-facing products to support end-user journeys or experiences (ordering, bill paying, and loyalty programs, for example). The other is around teams that build and manage the underlying platforms that product teams consume, such as customer relationship management (CRM) and marketing technology. These teams are made up of cross-functional people, with business taking a leading role. They operate using agile principles, bring on top talent to work on specific products and platforms, and have sufficient autonomy to develop solutions.

**Why this is important**

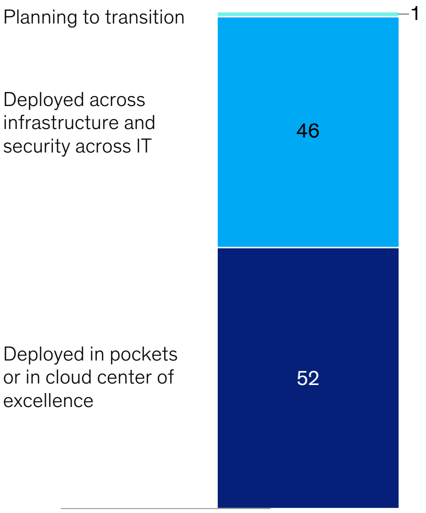
Cloud offers many speed and flexibility benefits, but it requires companies to change how they work to take advantage of them. Many institutions still use outdated, inefficient, and overly bureaucratic approaches. More damaging is when the business is not actively involved in the work and essentially "outsources" it to IT. This disconnect undermines IT's ability to harness cloud to generate value for the business. If organizations continue to work slowly, cloud's advantages in speed and flexibility are effectively lost.

For example, some infrastructure leaders have kept their outdated IT service-management processes (change-, release-, and incident-management processes), organizational silos (functional silos in infrastructure), and manual ticket-based processes even as they've moved applications to cloud, resulting in little or no benefit.

Exhibit 30

**Most surveyed firms have deployed an agile product model in pockets or within some teams.**

**Use of agile product model,[1]** % of companies (n = 67)



Note: Figures do not sum to 100%, because of rounding.
[1]Q: *Are you using an agile product model to enable your cloud program?*
Source: CloudSights

In contrast, infrastructure leaders who successfully transition to a more flexible product operating model that works at scale can see 30 to 40 percent improvements in labor productivity, 50 to 60 percent increases in resiliency, and 50 to 80 percent or more improvement in time to market.

**State of the industry**

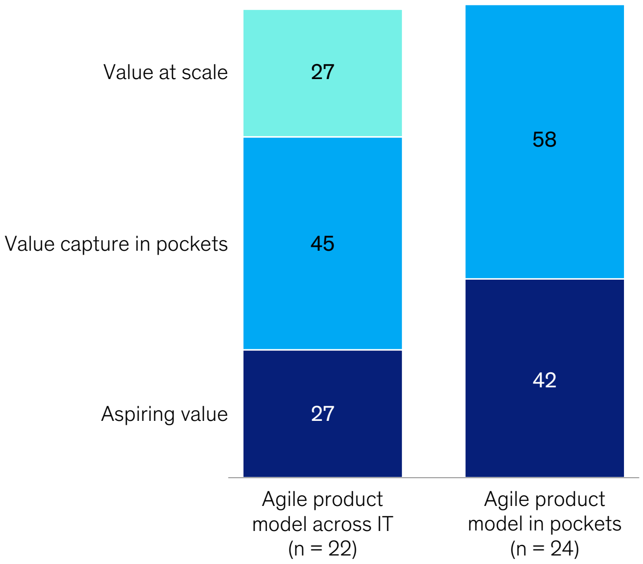
In our surveys, we found that more than 90 percent of organizations are already using some form of an agile product model to enable their cloud program. However, most of these organizations have only deployed the model in pockets or in their cloud center of excellence, rather than broadly across IT or even across the business (Exhibit 30).

Perhaps more important, organizations are seeing a correlation between agile adoption and cloud value realization. Those that have implemented agile across IT, rather than just in pockets, have realized more-substantial value (Exhibit 31).

Exhibit 31

## Value at scale clearly correlates with adoption of agile across IT.

**Cloud value realization,** % of companies (n = 46)



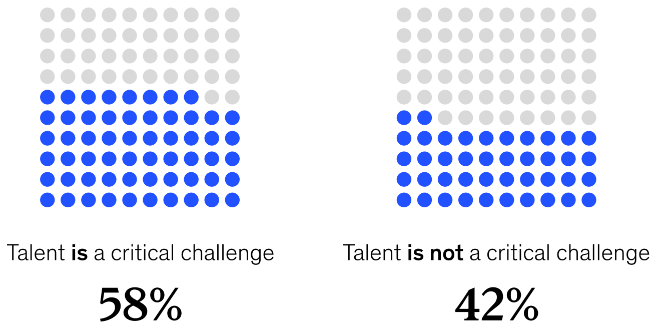
Note: Figures may not sum to 100%, because of rounding.
Source: CloudSights

Finally, organizations recognize the need for top talent, especially as they increase cloud adoption (Exhibit 32). Transitioning from outsourced system administration skill sets to insourced engineering skill sets is a critical enabler of cloud adoption.

Exhibit 32

## Most companies see talent as a top challenge as they increase cloud adoption.

**Talent criticality,**[1] % of companies, n = 31



Organizations with fewer (mostly horizontal) workloads on cloud found talent to be an inhibitor **only 50% of the time.**

**Talent became a top challenge** as organizations grew cloud adoption beyond horizontal workloads to vertical workloads.

The organizations for which talent is not a critical challenge **had high access to CSP engineers during the early days of adoption** or in the beginning of their cloud journey.

[1] Q: *Is lack of cloud talent a critical challenge for your organization?*
Source: CloudSights

### Components of success

— Implement an agile product and platform operating model for cloud built on six pillars (Exhibit 33):

1. **Build up engineering skills and attitudes.** Engage with app teams at their level of maturity while developing engineering-centric capabilities, such as CI/CD, modern ITSM processes (change, release, incident, and so on), and embedded cross-functional skills (including business, development, operations, infrastructure, risk, and security). Avoid inserting translators between IT and the business and instead move toward teams where IT and business leads are closely integrated.

2. **Create a great developer experience.** Unlock developer productivity through sharing and reuse of preconfigured tooling and application patterns made accessible through an easy-to-use developer portal. There should be a unified consumption layer with self-service for developers, as well as a standardized tech stack to support speed and agility with the highest degree of safety.

3. **Fully productize infrastructure services.** Treat infrastructure like a product. That means implementing automation, infrastructure as code, and self-service across the full life cycle (including “day-two operations”). This reduces manual ticket management, unplanned work, and siloed coordination. Cross-functional infrastructure teams have end-to-end responsibility for delivering and automating standardized infrastructure and platform products offered via APIs that cover both private and public cloud.

4. **Integrate security by design into development.** Distribute security governance, built-in cyberrisk management, and compliance checks throughout the SDLC, and support application teams with security talent in every component of the operating model. Security products need to be consumable.

5. **Recruit and retain top engineers by developing an engineering culture.** Top engineering talent can be up to ten times more productive than an average engineer. But to attract, grow, and retain top talent, companies need an engineering-first culture that offers autonomy, rapid iterations, no bureaucracy, reduction of low-value operational work, and opportunities for development.

6. **Install a governance model focused on outcomes.** The cloud platform team sets centralized governance standards that give teams the ability to develop apps independently, ensuring full autonomy within controls. Develop OKRs focused on business outcomes rather than technology activities (such as number of tickets resolved). Track progress against these OKRs, and ensure that all teams use them so that leadership has full transparency into what teams are doing and can hold them accountable.

— **Develop tailored engagement models to support teams.** In a typical large-enterprise IT organization, some application teams have highly mature capabilities, while others lack the skill sets to really take advantage of cloud. It’s therefore important to define a set of tailored “engagement models” between the central cloud platform team and product teams to ensure they know how to work together. A sample set of engagement models might include the following:

   » **Concierge:** full support by central cloud platform team for product teams with nascent skills and experience in cloud

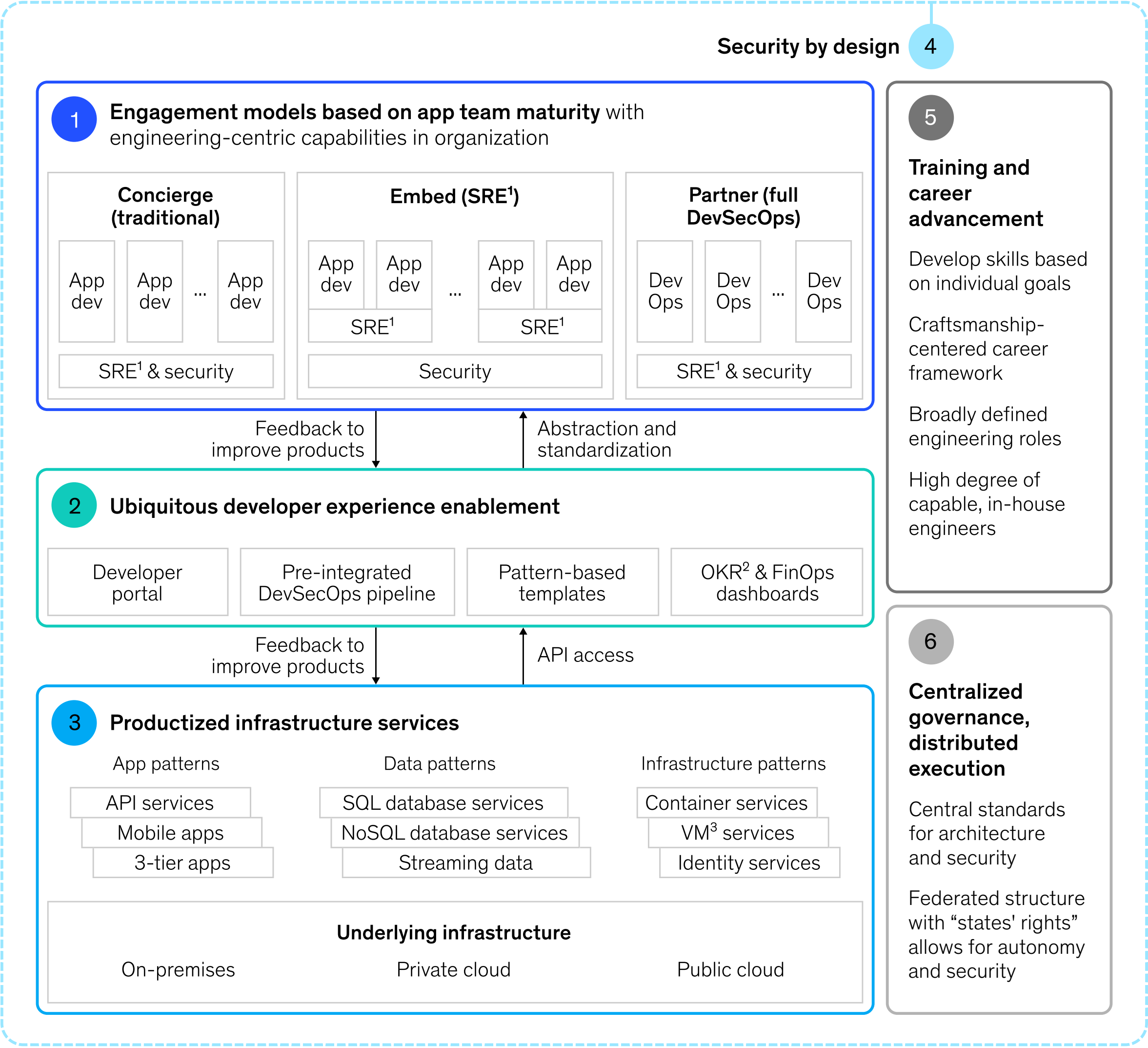
   » **Embedded:** central cloud platform team owns the cloud foundation base and key isolation or landing zone components, while mature product teams manage operational and security disciplines

   » **Partner:** product teams with the greatest level of maturity operate in a full “you build it, you run it” model with support as needed from a central cloud platform team

Exhibit 33

## An effective product operating model for cloud is built on six pillars.

**Illustrative**



Security by design
4
1
Engagement models based on app team maturity with engineering-centric capabilities in organization
Concierge (traditional)
App dev
App dev
...
App dev
SRE¹ & security
Embed (SRE¹)
App dev
App dev
SRE¹
...
App dev
App dev
SRE¹
Security
Partner (full DevSecOps)
Dev Ops
Dev Ops
...
Dev Ops
SRE¹ & security
Feedback to improve products
Abstraction and standardization
2
Ubiquitous developer experience enablement
Developer portal
Pre-integrated DevSecOps pipeline
Pattern-based templates
OKR² & FinOps dashboards
Feedback to improve products
API access
3
Productized infrastructure services
App patterns
API services
Mobile apps
3-tier apps
Data patterns
SQL database services
NoSQL database services
Streaming data
Infrastructure patterns
Container services
VM³ services
Identity services
Underlying infrastructure
On-premises
Private cloud
Public cloud
5
Training and career advancement
Develop skills based on individual goals
Craftsmanship-centered career framework
Broadly defined engineering roles
High degree of capable, in-house engineers
6
Centralized governance, distributed execution
Central standards for architecture and security
Federated structure with “states' rights” allows for autonomy and security

[1] Site-reliability engineering.
[2] Objectives and key results.
[3] Virtual machine.

A leading SaaS company transformed fully toward a DevSecOps/SRE model as part of its hybrid-cloud scale-up strategy. Early adoption of public cloud without moving its teams to the new operating model didn't allow the company to realize full benefits. As it scaled its cloud adoption, it transformed the technology organization front to back to take full advantage of cloud automation and make its teams significantly leaner.

The transformation of the teams moving to cloud resulted in a more than 65 percent increase in resiliency, more than 50 percent improvement in time to market, 20 to 30 percent more efficiencies in infrastructure/operations teams, and significant improvements in developer experience (80 percent of developers registered higher satisfaction scores).

## 9 Modernize applications to get the full benefit of cloud services

Workload migration and modernization approaches vary in intensity, ranging from simple lift and shift to modernizing apps through rehosting and rewriting. The difference in the approaches essentially determines whether an application can merely operate effectively in cloud or take full advantage of its services and capabilities.

**Why this is important**

Cloud adoption is a prerequisite for achieving value at scale, but it is not sufficient on its own. Lift and shift—that is, migrating applications without modernizing them in any way—often results in little benefit beyond exiting a costly data center and in applications that are more expensive to run in cloud than on-premises. In contrast, a migrate-and-modernize approach can generate immediate benefits in cost, efficiency, scalability, and resiliency while minimizing incremental investment.

The range of migration and modernization approaches can be described in terms of levels of intensity and value (Exhibit 34). Companies that get more value from cloud go beyond a basic migration approach (Levels 1 to 3) to modernization of applications (Levels 4 or 5). For example, a Level 3 migration can enable an application to take advantage of basic cloud capabilities, such as scaling and automated security controls, while a Level 5 modernization can accelerate the path to value and take advantage of more cloud services. Companies should consider skipping migration Levels 1 to 3 and focus on Levels 4 or 5 for high-potential applications.

**State of the industry**

Organizations have used a variety of approaches to migrate and modernize their applications (Exhibit 35). Basic lift and shift (no application modernization, Level 1) has been the most common, especially when exiting data centers is the priority. Lift and optimize (basic optimizations, Level 2) is the next-most-common approach.

**Components of success**

— **Prioritize workloads to modernize.** Evaluate the feasibility of workload modernizations and their expected value to decide which ones to prioritize. Our recommendation is generally to do a minimal level of infrastructure modernization and then turn to applications that have the greatest business value potential.

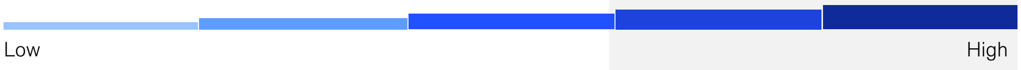
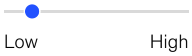
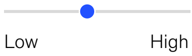
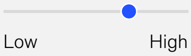
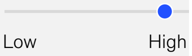
— **Start with a lighthouse modernization.** Take a high-value workload that is well suited for a Level 4 or 5 modernization, and work it through the process. This is generally much more successful than attempting multiple modernizations at the same time.

— **Modernize at scale.** There are two primary elements to focus on in order to scale modernization programs. One is to use application patterns, which are codified capabilities that can be reused across applications that have similar architectural characteristics. Workloads will often have similar architectural archetypes. Second is to establish a “modernization factory,” essentially a dedicated team to develop and apply application patterns across the workload pipeline, containerize workloads, verify functionality, and codify processes, among other tasks. Evaluate services provided by CSPs to determine if any can be helpful in the modernization process.

Exhibit 34

## Unlocking value in cloud requires more advanced levels of app modernization.

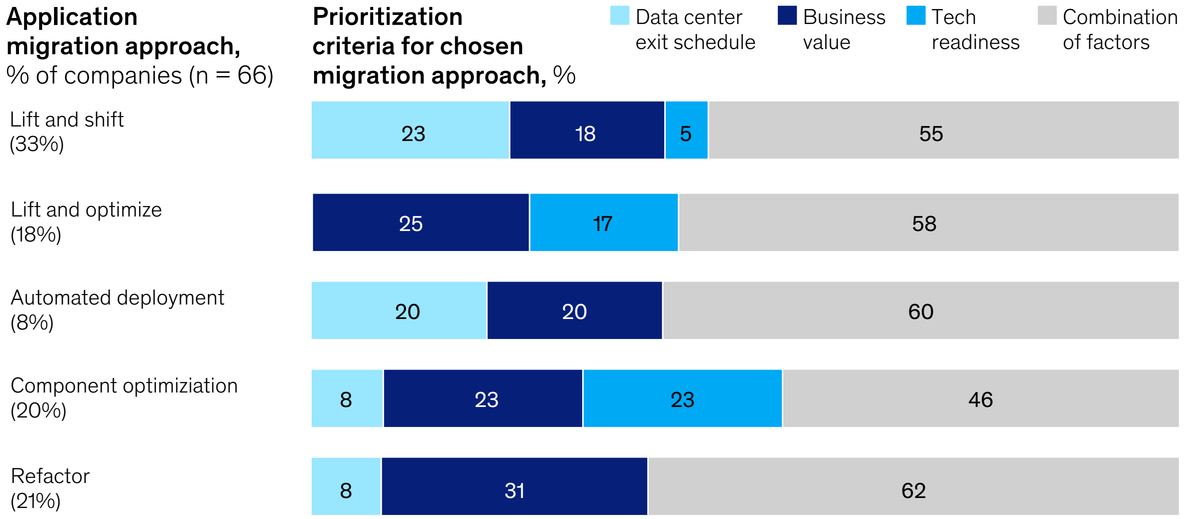
▨ Modernization, not migration

| | Enter the platform, improve tech speed and productivity benefits | | Application performance and agility improvement | | |
|---|---|---|---|---|---|
| | Manual deployment | Some automated deployment by infrastructure team | VM-based[1] automated deployment | Containerized | Serverless |
| | **Level 1: Lift and shift** | **Level 2: Lift and optimize** | **Level 3: Replatform; minimal preoptimization** | **Level 4: Rehost/rearchitect** | **Level 5: Rewrite** |
| **Return on investment** |  Low | | | | High |
| **IT cost savings** | **App maintenance productivity** | 0% | 20% | 25% | 30% |
| | **Infrastructure productivity** | 5% | 20–30% | 30–40% | 40–50% |
| | **Hardware cost savings** | 10% | 30% | 40% | 50% |
| | **Operating expenditure cost savings** |  Low High |  Low High |  Low High |  Low High |
| **Speed and agility** | Limited or no improvement (eg, infrastructure provisioning still manual) | Moderate improvements in infrastructure provisioning (eg, faster ticket turnaround time through IaC[2]) | Quicker deployment (adoption of enterprise CI/CD[3] pipeline for both applications and infrastructure | More sophisticated experimentation (eg, A/B testing, test and learn) | Faster time-to-value (developer productivity through modularity and abstraction) |
| **Risk and security** | Existing security debt/posture migrates to cloud | Low risk with automated infrastructure controls and limited automated app security controls (ie, code scanning, tokenization of sensitive data) | Reduced risk with automated app and infrastructure security controls as well as automated compliance pre- and post-deployment | Modern containers reduce threat vectors, while automated tooling improves security | Using managed services from CSP shifts infrastructure security responsibility to cloud provider |
| **Decision rationale** | Lift-and-shift migration when time and investments are very limited (eg, upcoming deadlines for data center exit) | Infrastructure-led migration with minimal dependency on application teams | Optimized migration to unlock agility benefits and automated controls while minimizing investments | More advanced migration for select applications (eg, when containerizing is easy) | Used for new applications built directly on cloud using native services |
| | | | Optimal level of **migration** to achieve value from cloud | | Optimal level of **modernization** to unlock full value from cloud |

[1] Virtual machine.
[2] Infrastructure as code.
[3] Continuous integration / continuous delivery.

Exhibit 35

## Lift and shift is the primary migration approach, though respondents focused on business value choose a variety of approaches.



Note: Figures may not sum to 100%, because of rounding.
Source: CloudSights

- **Use generative AI tools to transform modernization economics.** Generative AI–based tools can make moving applications to cloud faster and less expensive by allowing developers to spend less time analyzing legacy code bases, writing boilerplate code, and integrating platform services, but only if they are applied correctly (table). No generative AI migration tool should be a black box. It should empower an engineer through automated dependency mapping, automated documentation of program structure, and suggestions for object-oriented classes and functions required to create a modern code base. Early experiments in applying generative AI to application remediation suggest that time and cost can be reduced by 50 percent.

- **Enforce holistic governance.** A holistic governance process that ensures business alignment and value capture should evaluate performance indicators throughout the modernization process.

## 10 Use FinOps to control and optimize cloud spend

FinOps (financial operations) is a cross-functional collaboration between engineering, platform, finance, and the business to establish capabilities to track, allocate, and optimize cloud spend. A mature FinOps capability can provide a detailed view of the economics of cloud usage (compute, storage, network) to help plan for cloud costs and manage them on an ongoing basis. As companies embark on the next cycle of innovation through adoption of new "cloud-hungry" technologies such as generative AI, it is paramount that they invest in improving their FinOps maturity, which can make the difference between a positive and negative business case.

Table

## Generative AI tools provide opportunities to accelerate and improve cloud-related tasks.

**Example of generative AI capabilities that can be used across different categories of tasks**

| Data | Code | Cross-cutting |
|---|---|---|
| **Aggregate synthesis of business domains from code base**<br>Iteratively extract and model connections between components<br>**Requirements synthesis**<br>Create component-level English description of requirements through domain expertise and prompt engineering<br>**Component algorithmic flow synthesis**<br>Describe key formulas and expected behavior of output of code file<br>**Data structure synthesis**<br>Extract structure of legacy data, consolidate references, and create target database structure<br>**Cross-interface synthetic bounded context**<br>Synthesize ingestion of SQL and stored procedures into English description through generative AI agents | **Automated code documentation**<br>Auto-generate high-level documentation that synthesizes software definition<br>**Requirement-to-component synthesis**<br>Map high-level requirements to code files that implement those requirements<br>**Cross-component memory context**<br>Preserve context of code generation across multiple code files<br>**Context window management**<br>Exclude memory of certain code context based on requirements mapping to code through generative AI agents<br>**Target-state code creation**<br>Generate target-state code (eg, Python, Java, .NET, etc) that mimics the functionality of the initial component with a developer lens<br>**Business-rule extraction**<br>Mine business rules embedded in the legacy source code<br>**Code pattern recognition**<br>Automate identification of repeated code patterns within the portfolio of applications | **Developer experience**<br>Expedite, automate, and simplify heavily manual and monotonous activities<br>**Optimizations**<br>Identify data and compute opportunities to use distributed parallel execution<br>**Target-state architecture deployment**<br>Target code generation into specified target-state architecture<br>**Unit test generation**<br>Build test cases for different components descriptively<br>**Observability data incorporation**<br>Use incident/utilization logs to generate dynamic calls |

**Why this is important**

While cost savings should not be the primary reason for going to cloud, it is true that doing so can help organizations realize meaningful efficiencies. Many, however, find themselves stuck in a spiral of continuously growing costs without much clarity on what is causing them, and they begin to wonder about the wisdom of the whole initiative.

In some cases, for example, cloud can actually cost more, unit for unit, than an on-premises system, especially when infrastructure has been fully depreciated. Cloud comes with a set of value-added services that add to unit costs, such as out-of-the box resilience, self-service capabilities, and others.

The real cost savings come from cloud's elasticity—its ability to scale up and down as needed—which allows the use of exactly the amount of capacity needed and no more. But this advantage only holds if an organization really limits itself to exactly what is needed, which requires careful attention. Organizations whose FinOps capabilities are mature can reduce cloud spend by as much as 20 to 30 percent—and in many cases, about half of those savings can be achieved quickly.[15]

Generative AI dramatically increases the importance of FinOps (and FinOps as code) because it accelerates and changes the shape of cloud consumption. The rapid evolution and enterprise adoption of generative AI by businesses is likely to be a disruption to cloud programs in multiple ways—for example, increasing proliferation of use cases that drive higher cloud consumption, integration with more hyperscalers to implement specific use cases, and accelerating migration with the aid of generative AI enablers.

Moreover, increased adoption of  generative AI could dramatically shift costs toward data management. So far, cloud costs have been predominantly compute costs, with data costs following at a distant second. With generative AI, large data sets are needed to train and refine models, with per-terabyte charges, especially when using audio and video rather than simply text. This increases data costs substantially, driving overall cloud costs higher and upending the relationship between storage and compute costs. Once generative AI models reach production, companies must consume tokens to process each prompt that users enter. If the model is not fine-tuned properly, users entering repeated, complicated prompts will lead to skyrocketing token consumption costs.

### State of the industry

Many organizations are still in the early stages of building their FinOps capabilities and have yet to fully realize the benefits of this approach, according to our research (Exhibit 36). One of the reasons reported in a McKinsey survey is that many CFOs, chief procurement officers, and business unit heads don't become meaningfully involved in cloud programs until annual costs surpass $100 million. As a result, they miss earlier opportunities to realize significant savings.

In the early stage of maturity, many teams are focused on making costs visible (through a show-back model, for example). This goal is greatly complicated by the fact that the tooling market is still quite fragmented, with no single tool providing all the FinOps capabilities an organization needs. Almost three-quarters of companies rely on third-party tools for their financial analysis (Exhibit 37).

### Components of success

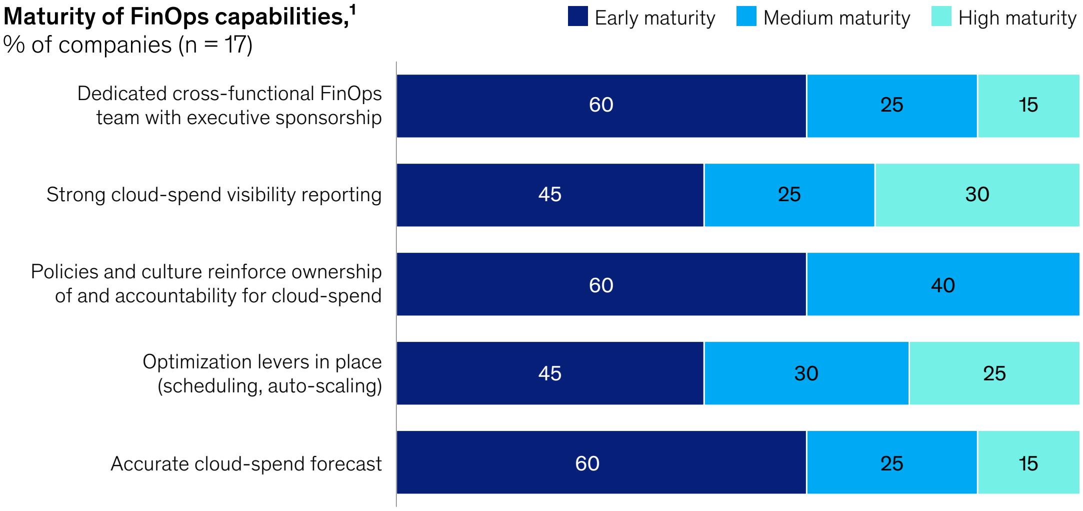
— **Use FinOps from the outset.** Institutions embarking on a cloud journey often must weigh a multitude of competing priorities, and it is easy for FinOps to get lost in the shuffle. But the longer the delay, the more difficult it will be to implement FinOps, given the increased complexity of environments and proliferation of suboptimal provisioning practices.

— **Involve business leaders early in the process.** Business leaders tend to get involved with cloud programs only after its costs become significant. By this time, however, IT has generally made substantial moves in cloud, and unwinding or modifying them can be time consuming and expensive. For this reason, it's crucial to involve business leaders early in cloud-migration efforts by establishing joint accountability and providing clear cost–benefit analysis of app performance on cloud.

[15] Keith Conway, Abdallah Saleme, Bhargs Srivasthan, and Konstantin Tyrman, "The FinOps way: How to avoid the pitfalls to realizing cloud's value," McKinsey, January 18, 2023.

Exhibit 36

## FinOps capabilities are at an early stage of maturity.



[1] Q: *How would you currently rate your FinOps capabilities?* Respondents could select more than one option.
Source: McKinsey FinOps survey

- **Help teams optimize their cloud spend.** In addition to operational activities like tagging and reporting, FinOps teams need to work with cloud teams to develop charge-back or show-back mechanisms for transparency and accountability.

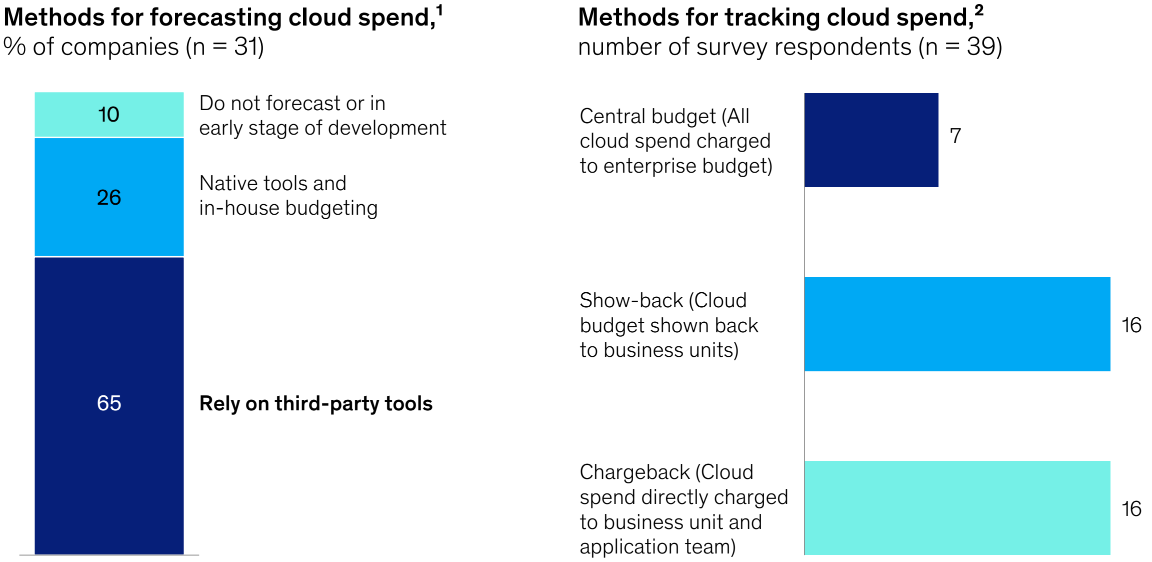
- **Ensure wide adoption of FinOps practices within engineering communities.** FinOps programs should be accompanied by appropriate change-management actions that include incentives, community building, and communication. The goal is to effectively “shift left” so that a FinOps perspective on cloud economics options is integrated into the operating model before development starts.

- **Introduce FinOps-as-code capabilities.** This relatively novel concept refers to the automation of financial-management principals as part of the software development life cycle. FinOps as code (FaC) employs a combination of automation, tooling, and cloud-native services to identify meaningful cloud cost insights before code is deployed. Additionally, cost controls can be defined to set and enforce budgets, identify areas of cost reduction, and execute actions like scaling down nonessential resources. FaC can be applied to both new and existing applications and infrastructure. Companies can start by defining cost standards for infrastructure provisioning, embedding them into application patterns, setting up budget alerts when costs exceed expectations, and deploying simple rules to reduce waste, such as releasing unattached disk volumes that store no data.

- **Deepen the understanding of cloud unit economics.** One of the most important capabilities that FinOps can provide is showing a clear relationship between cloud consumption costs and business value.[16] This allows the technology side to talk to the business side “in their language” to drive decisions together as one cross-functional team.

Exhibit 37

## Most companies rely on third parties to forecast cloud costs.



Note: Figures do not sum to 100%, because of rounding.
[1]Q: *How do you forecast cloud spend?*
[2]Q: *How does your organization do cost tracking today? What is the visibility and impact of cloud spending on individual app teams?*
Source: CloudSights

— **Apply three lines of defense to help keep generative AI–driven disruptions under control:**

1. **Upstream:** Create visibility into the unit cost for each use case, helping validate that each one has a positive ROI (thereby also reducing the number of use cases by eliminating those that aren't worth it).

2. **Midstream:** Apply a set of levers to optimize both the token price (for example, by using an open-source or a lower-tier version of the model) and the quantity of tokens consumed (for example, by applying effective prompt engineering).[17]

3. **Downstream:** Create the telemetry to monitor and continuously report on the overall cost and unit-economics performance of generative AI models. This can help organizations shut down costly usage, retune inefficient models, and ensure that the use cases that are not profitable are discontinued.

Generative AI offers additional capabilities to FinOps as well, from better descriptive presentation of the data to root-cause analysis, reducing strains on FinOps analysts. One large global life sciences company is using generative AI as a virtual FinOps analyst, allowing developers and product owners to ask questions about their spending patterns—for example, "What cloud services have increased the most over the past month?" or "What does my application id=1500 cost per month for storage, by storage type?"—and generate real-time dashboards

[16] "The FinOps way," January 18, 2023.
[17] In the realm of natural-language processing, prompts are intrinsically tied to free-form text, and their input-output dynamics are quantified in terms of tokens, where a token typically represents a four-letter portion of the word.

with answers. The generative AI–trained LLM then automatically charts answers into an interactive dashboard on demand. This reduces the load on the central FinOps team to field these types of questions and puts the power of real-time answers closer to the consumer.

A major US retailer that was an at-scale cloud consumer invested in FinOps capabilities and focused on short-term savings, with an eye toward sustainable capabilities. The team is composed of people from finance, data analytics, architecture, and engineering, federating FinOps across the cloud consumers in its enterprise. A short-term focus on FinOps helped the company save 20 percent on its annual cloud consumption and put it in a position to scale the capabilities going forward. Home-grown data analytics and dashboarding resources make cloud spend more visible and digestible. This level of scrutiny saved 24 percent on cloud last year and even greater savings are anticipated this year.

Cost tracking and funding is centrally managed but anything directly attributable is charged back to the relevant business unit. For example, the data warehouse and data lake are hosted in cloud, and the central team tracks consumption by business user. This is charged back to the business, while the setup and migration is done centrally.

❖ ❖ ❖

Cloud is rapidly shifting from a business enabler to a business necessity. Harnessing modern technologies such as generative AI, operating with speed and flexibility, and continuously improving and adapting while managing risk are the hallmarks of a successful business in the age of digital and AI. They are also increasingly impossible to develop without having cloud at the core of operations.

**Chhavi Arora** is a partner in McKinsey's Seattle office, **Will Forrest** is a senior partner in the Chicago office**, Mark Gu** is a partner in the Boston office, and **James Kaplan** is a partner in the New York office.

The authors wish to thank the following people for their contributions to this report: Ritesh Agarwal, Bridget Ahad, Noorul Akram, Will Ardron, Aaron Bawcom, Jan Shelly Brown, Bailey Caldwell, Charles Cao, Bill Gregg, Colin Gunter, Ashish Gupta, Rich Isenberg, RJ Jafarkhani, Kaz Kazmier, Venkatesh L, Wasim Lala, Ling Lau, Ani Majumder, Anand Mohanrangan, Chett Rubenstein, Kaavini Takkar, Konstantin Tyrman, and Fayet Valles.

McKinsey Digital
November 2023

Designed by Darby

www.mckinsey.com

@McKinsey
@McKinsey