run:ai

The 2021 State of AI Infrastructure Survey

October 2021

Table of Contents

Introduction and Key Findings

Introduction

Most research around the state of the AI industry talks about the same few facts. AI is still very immature, models rarely make it to production, and there are a lot of challenges for data scientists and research teams around creating the right infrastructure and setting up AI for success.

To get more insight into whether these pervasive ideas are still gospel in 2021, we commissioned a survey of 211 data scientists, AI, Machine Learning, IT and System Architects from 10 countries around the world. We spoke to people in the US, UK, Ireland, France, Spain, Italy, Germany, Finland, Sweden, Poland and Russia from companies of all sizes, many with over 5,000 employees, and some with as many as 10,000. We asked these enterprises to open up about the technologies they use, the challenges they face and the size of not only their budget, but also their confidence in their underlying abilities. The survey was completed by independent research company, Global Surveyz and the responses took place during June and July 2021.

The results are a fascinating look at the true state of AI maturity. We are working in a market with enormous potential. Three-quarters of those surveyed are looking to expand their AI infrastructure, and 38% have more than $1 million in budget per year to make that happen. However, big challenges definitely exist, some of which are very early-stage in terms of resource allocation, data usage, and goal setting. With so much invested in making AI a success, and companies looking to forge ahead and make progress, it's clear that early adopters of the right technology have a lot to gain.

Key findings

1 AI is a cloud-native world

AI was clearly born with the cloud in mind, with 81% of companies working cloud-natively (using containers) already and an additional Y% with plans in place to do so. Along with the use of containers comes adoption of Kubernetes and other cloud-native tools for management of containers. 42% are already using Kubernetes, another 13% on OpenShift, and 2% on Rancher. These numbers are considerably greater than container adoption for non-AI workloads, making AI a leader in cloud-native adoption.

2 Big spenders, but a lack of confidence

Our study shows that 33% of companies have a budget of more than $1M a year for AI infrastructure alone, and 59% have more than $250k a year. These huge budgets should in theory be providing some peace of mind that companies surveyed can get AI models into production. However, our survey found that for 77% of companies, less than half of models make it to production. 88% of companies say that they are not fully confident in their AI infrastructure set-up, and aren't sure that they can move their models to production in the timeline and budget provided.

3 Infrastructure challenges weigh heavily on AI teams

Lack of confidence in AI infrastructure extends to hardware utilization, with 83% of surveyed companies not fully utilizing their GPU and AI hardware, and 83% of companies admitting to idle resources or only moderate utilization. Only 27% say that GPUs can be accessed on demand by their research teams as needed, with almost half of those who responded relying on manual requests.

4 AI is still a relatively immature market

The top challenges for today's AI teams are data collection (61%), infrastructure/compute (42%) and defining business goals (36%). All three of the biggest challenges are early-stage problems for teams working with AI, which speaks to market immaturity. In addition, tools used to manage infrastructure for AI teams include home-grown tools (23%) and even Excel spreadsheets (16%), again showing that in many ways AI is still lacking maturity.

5 Budgets are growing, despite challenges

AI challenges are relevant across all respondents, regardless of company size, industry, AI spend, or infrastructure location (cloud, hybrid, or on-premises). Infrastructure utilization is an issue for between 85%-90% of respondents even among companies that have $10M or more budgeted for AI each year. Despite this many/most companies are not limiting their budgets until their challenges are solved, with 74% planning to increase spend on AI infrastructure in the next year.

6 AI has enormous potential for those who beat the challenges

There is strong pressure on enterprises to launch AI projects and to see value from Artificial Intelligence. While the challenges may still be early-stage issues like goal-setting and infrastructure set-up, the spend is far from immature. The financial support is in place to make a go of AI projects, but it needs to be channelled to the right places, improving the systems used for AI infrastructure management, solving hardware utilization challenges, and supporting research teams in gaining both confidence and access to resources.

Large Teams and Big Budgets Don’t Protect from Hardware Utilization Issues

GPU Farm Size and Server Locations

Over half of surveyed companies (53%) have GPU farms of 10 or more GPUs (figure 1), two-thirds (64%) are hosting their GPU in the cloud or hybrid and a third are running on-prem (figure 2). Over half (53%) already have their AI applications and infrastructure in the cloud, with another third (34%) planning to move to the cloud in the coming years (figure 3).

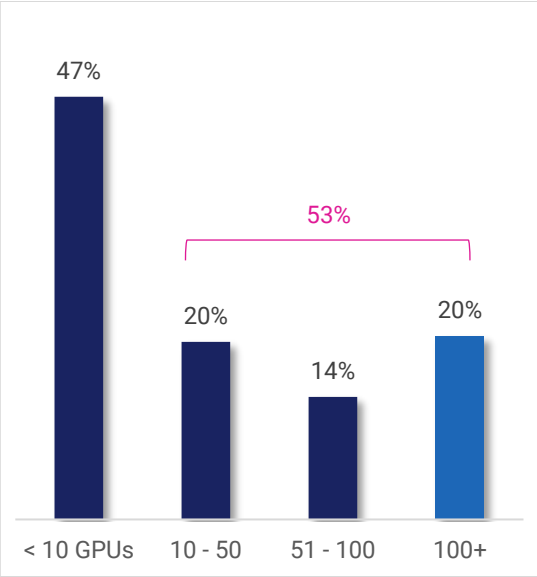

Figure 1 Size of GPU Farm

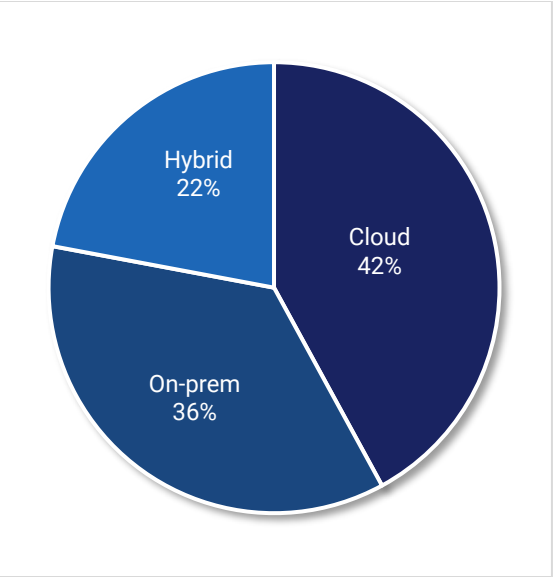

Figure 2 GPU Servers' Location

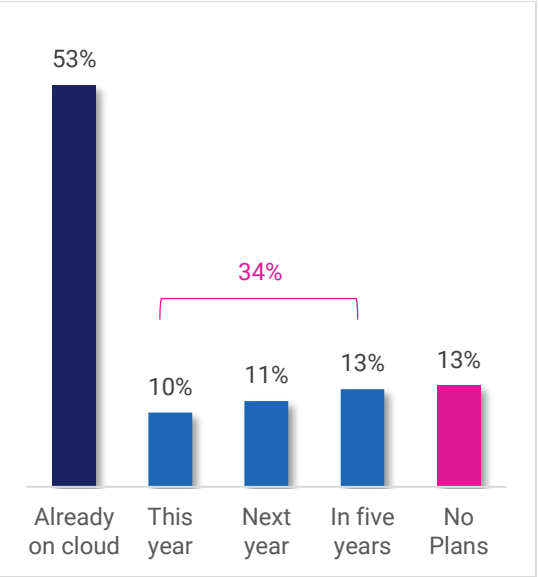

Figure 3 Plans for Moving AI Applications and Infrastructure to the Cloud

Size of Research Teams and Access to On-Demand GPU Compute as Needed

Almost two-thirds (63%) of companies have research teams of 10 or more and yet only 27% of them have solved the issue of fully on-demand access to GPU compute. The size of the research team doesn't equate with access capabilities.

Over a third (35%) do not have access to on-demand GPU compute, and almost half of this group (43%) require manual requests to gain access to GPU compute (figure 3). Every time they want to run a job, they need to make this manual request, slowing down operations significantly and adding a lot of frustration and delay.

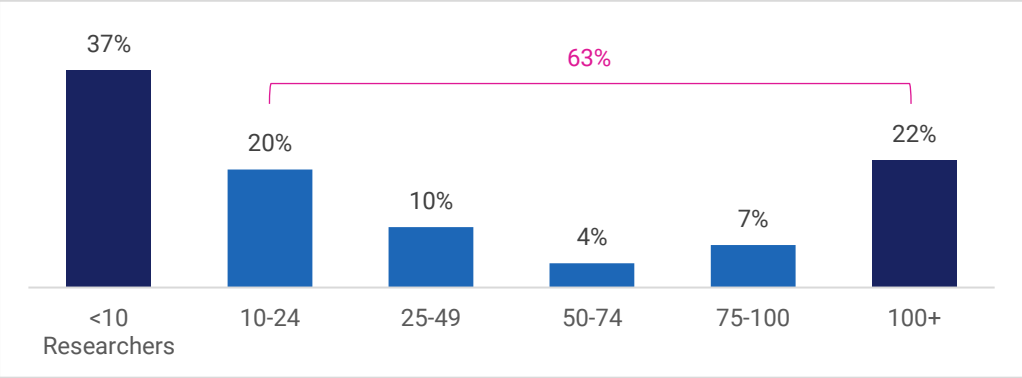

Figure 4 Size of Deep Learning Research Team

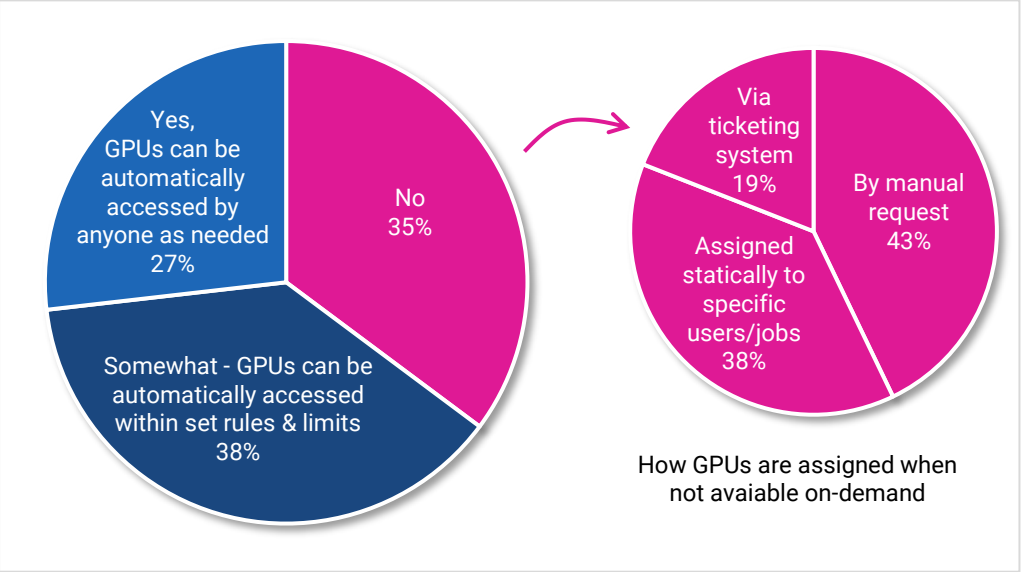

Figure 5 Do Research Teams Have On-demand Access to GPU Compute?

87% of respondents said they experience some level of GPU/compute resource allocation issues, with 12% saying this happens often. As a result, **83% of surveyed companies are not fully utilizing their GPU and AI hardware.** In fact, almost two-thirds (61%) indicated their GPU and AI hardware are mostly at moderate utilization (figure 6).

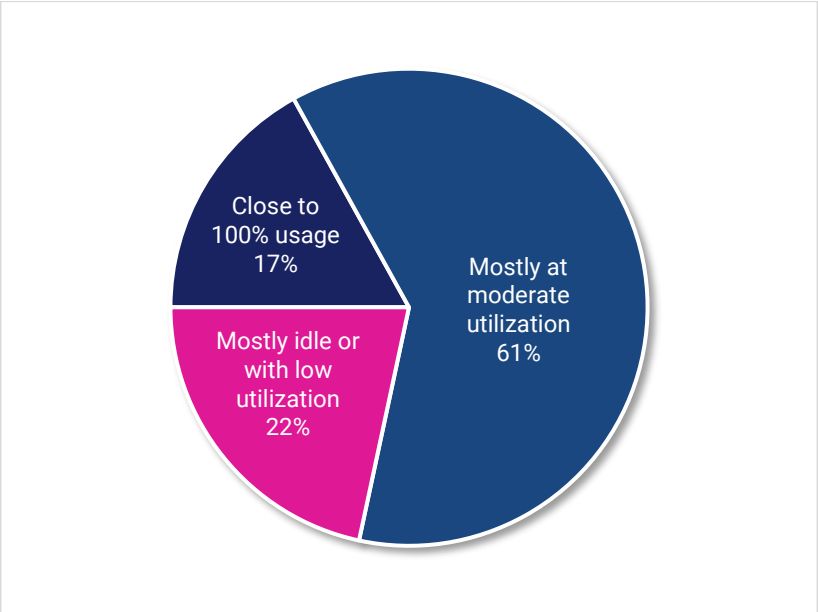

Figure 6 GPU and AI Hardware Utilization

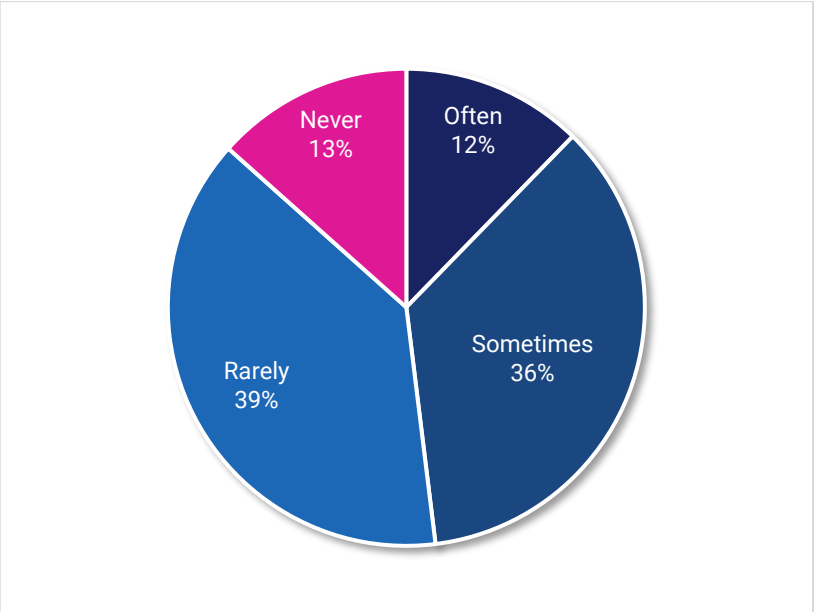

Figure 7 Frequency of Experiencing GPU/compute Resource Allocation Issues

Companies of All Sizes Struggle with Hardware Utilization

33% of companies have an annual AI infrastructure budget of over $1 million (figure 8).

When comparing budgets by level of AI hardware utilization, we see the companies with the smaller budgets of up to $250k suffer the most from having their hardware being mostly idle.

However, companies, at almost all budget ranges suffer from moderate GPU and AI hardware utilization. Only companies with the largest budgets of over $10 million were able to make a leap where 45% are able to get to close to 100% utilization. Even then, that still leaves 55% of this category struggling with moderate or mostly idle utilization.

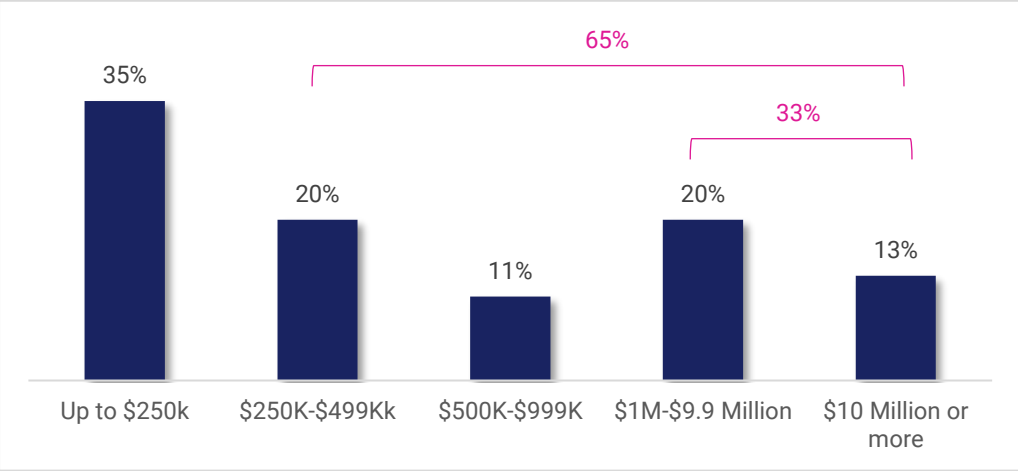

Figure 8 Annual AI infrastructure Budget (Hardware, Software, Cloud)

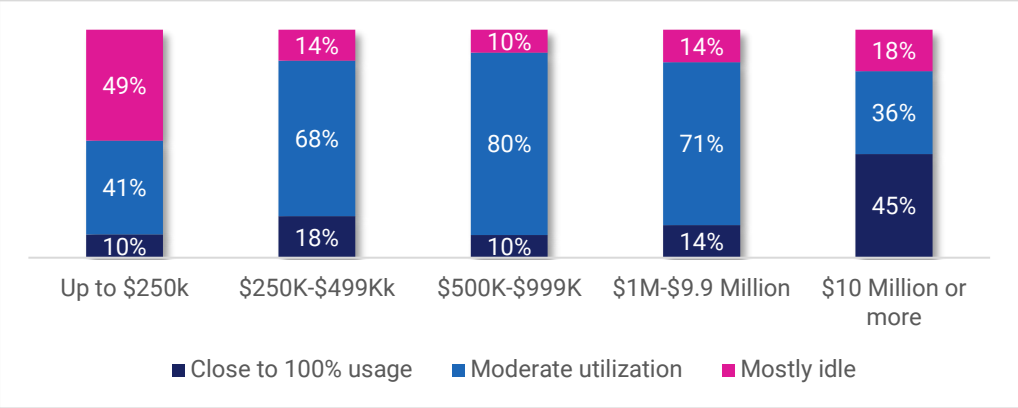

Figure 9 Annual AI infrastructure Budget by AI Hardware Utilization

Tools Used to Optimize GPU Allocation Between Users

72% of companies are using different tools to optimize their GPU allocation between users. From home-grown solutions (23%), to Excel spreadsheets (16%), a huge number are relying on low-tech solutions, especially when you consider the amount of budget that is being channeled into these projects.

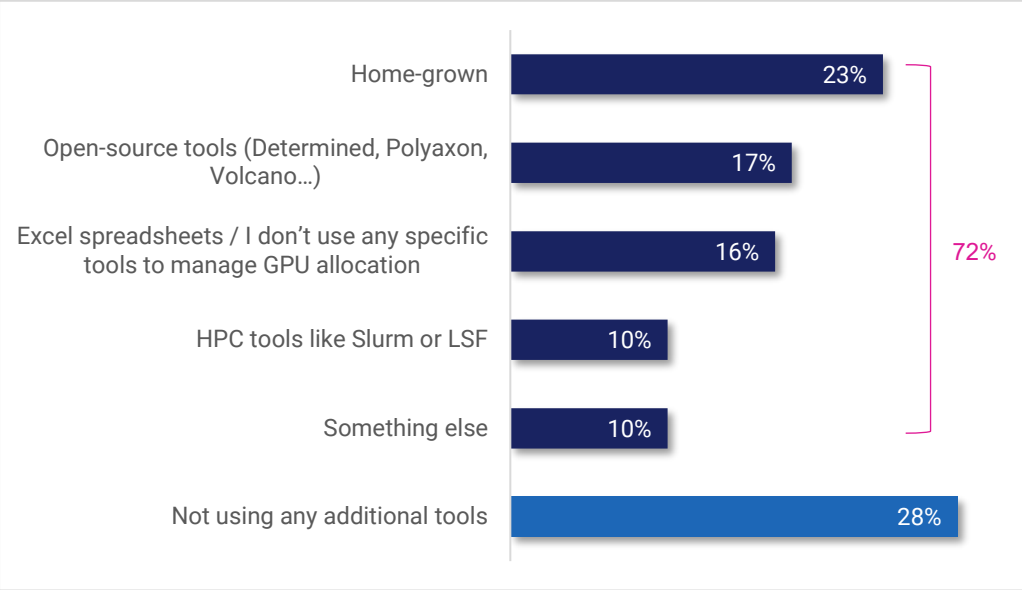

Figure 10 Tools Used to Optimize GPU Allocation Between Users

Containers and Kubernetes for AI Workloads

81% of companies are using containers for their AI workloads (figure 11) with Kubernetes ranking as the #1 container orchestration system, used by 42% of companies (figure 12).

These numbers show that AI is born in cloud-native infrastructure, and has a far greater adoption of cloud than the broader software world. Kubernetes is also ubiquitous with AI, with companies either using Kubernetes directly, or leveraging managed k8s through a third-party. The use of orchestration tools also shows that companies are confident and mature in their use of containers.

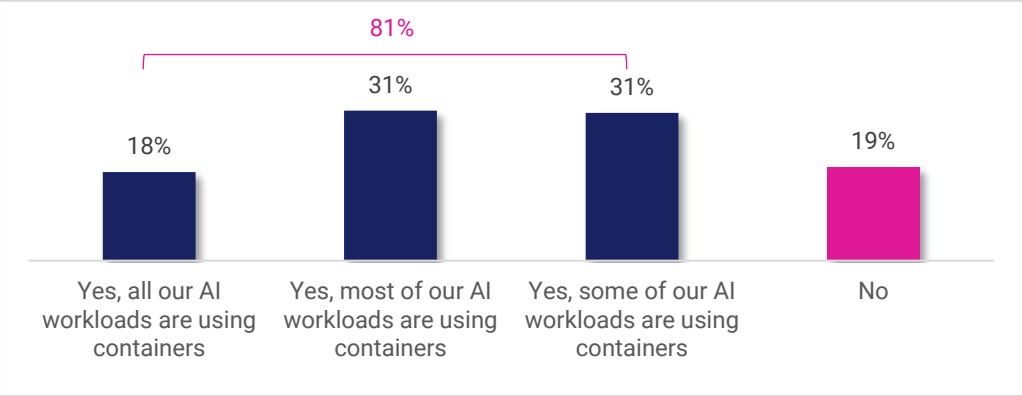

Figure 11 Use of Containers for AI Workloads

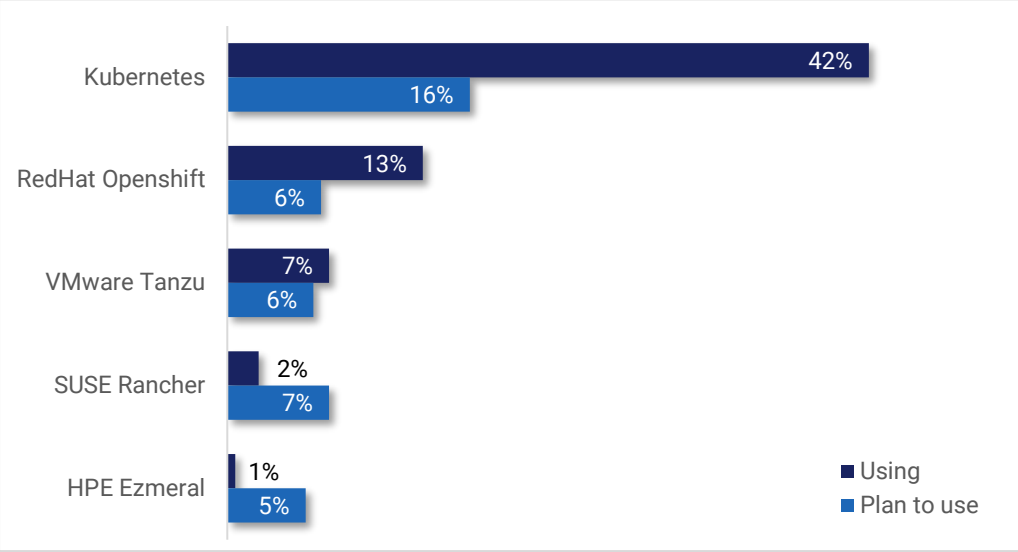

Figure 12 Container Orchestration Tools Used for AI Workloads

Big Plans for AI, Despite Multiple Challenges and Limited Confidence

Models Making it to Production

Less than half of AI models make it to production for 77% of surveyed companies.

Only 10% said 90% of their AI models make it to production.

Of course, often these AI models are experiments and it stands to reason they won't make it to production, but these numbers are still very high.

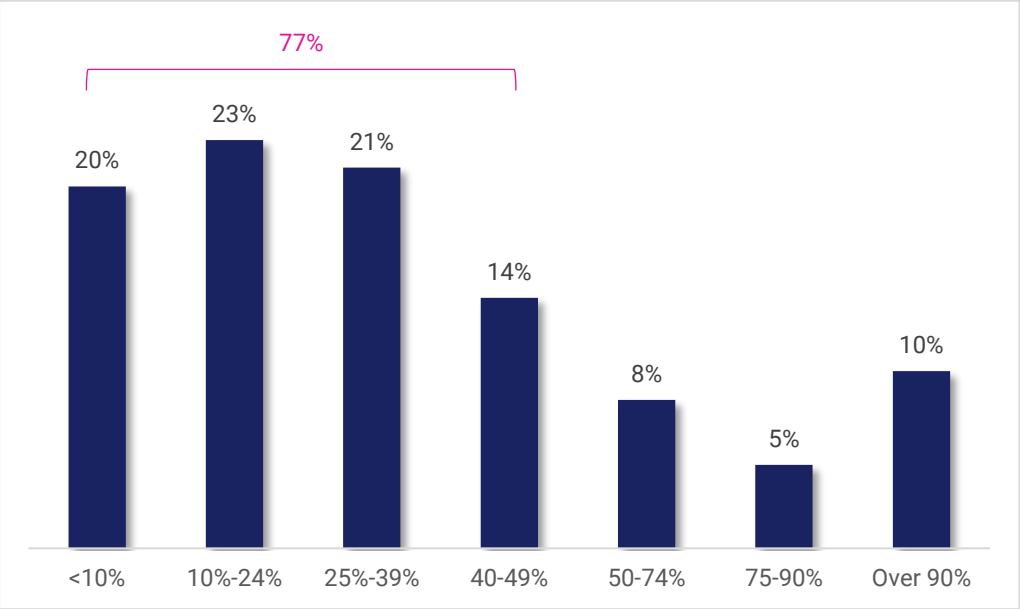

Figure 13 Models Making it to Production

Main Challenges for AI Development

96% of companies admit to having challenges when it comes to AI development.

The top three challenges are data-related challenges (61%), Infrastructure/Compute related challenges (42%), and defining business goals (36%).

These challenges are very immature, showing that companies are still working out how to get started with AI. Despite the fact that many companies have a $1M budget in place, they still aren't necessarily sure what their goals are or how to collect suitable data.

Infrastructure set up is a significant challenge, as companies struggle with visibility and control.

In general, the larger the company size, the greater the challenges become.

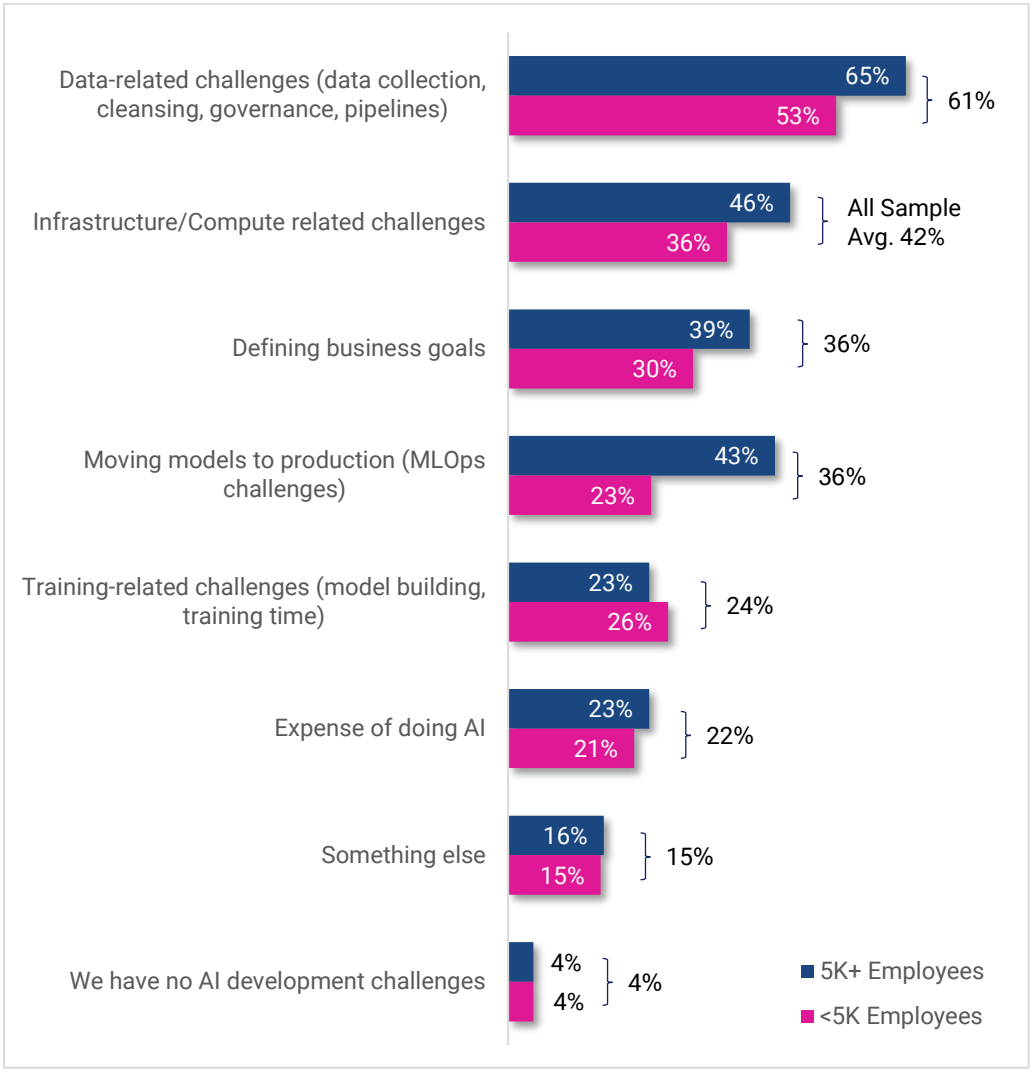

Figure 14 Main Challenges for AI Development

Plans to Increase GPU Capacity or Additional AI Infrastructure

Even with all of these challenges in place, 74% of companies are planning to increase their GPU capacity or AI infrastructure. Companies are confident in the growth of AI and are planning to increase capacity and spend, but will need to address their challenges to see success with their plans.

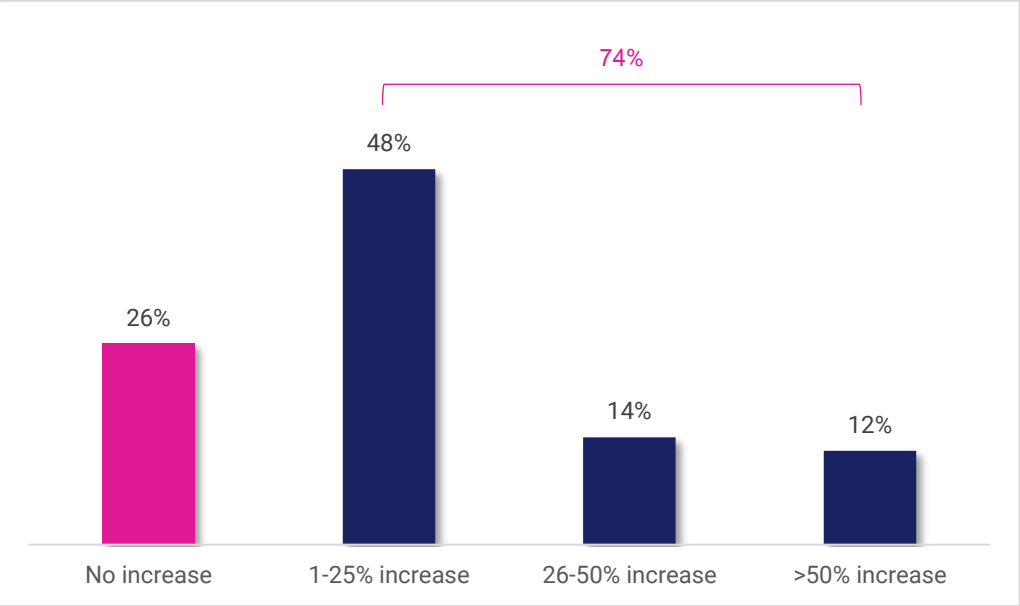

Figure 15 Plans to Increase GPU Capacity or Additional AI Infrastructure

Confidence in AI infrastructure Stack Set-up to Build, Train and Move

Only 18% of companies are fully confident that they have the right AI infrastructure stack set up to efficiently build, train and move ML models to production on time and on budget.

Despite a lack of confidence, and multiple challenges – companies clearly feel a pressure to keep moving, investing more budget and expanding AI plans to keep up with the competition.

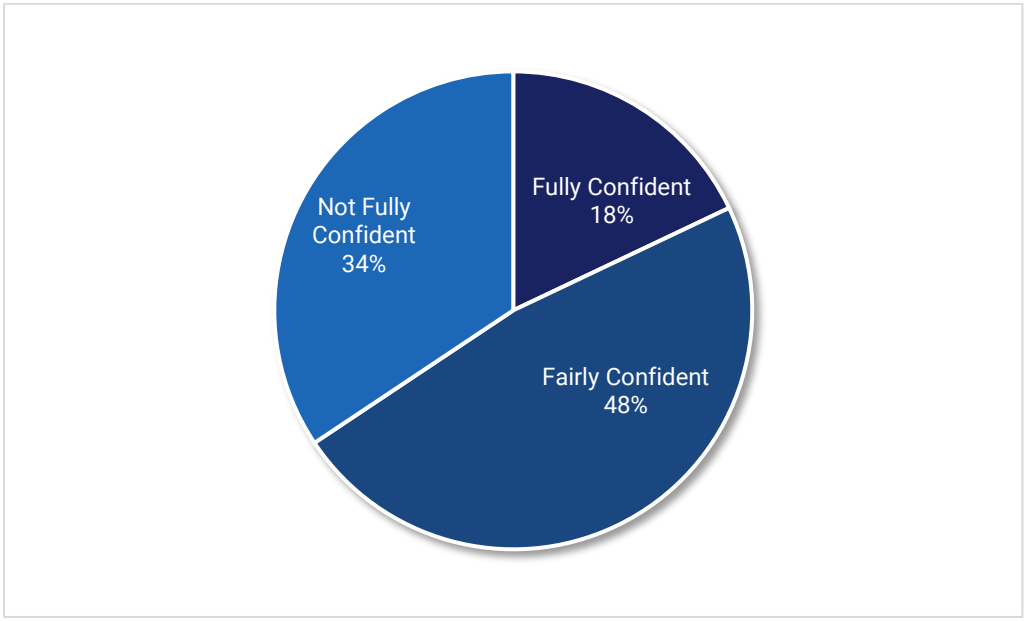

Figure 16 Confidence in AI infrastructure Stack Set-up

Demographics

Country of Residence

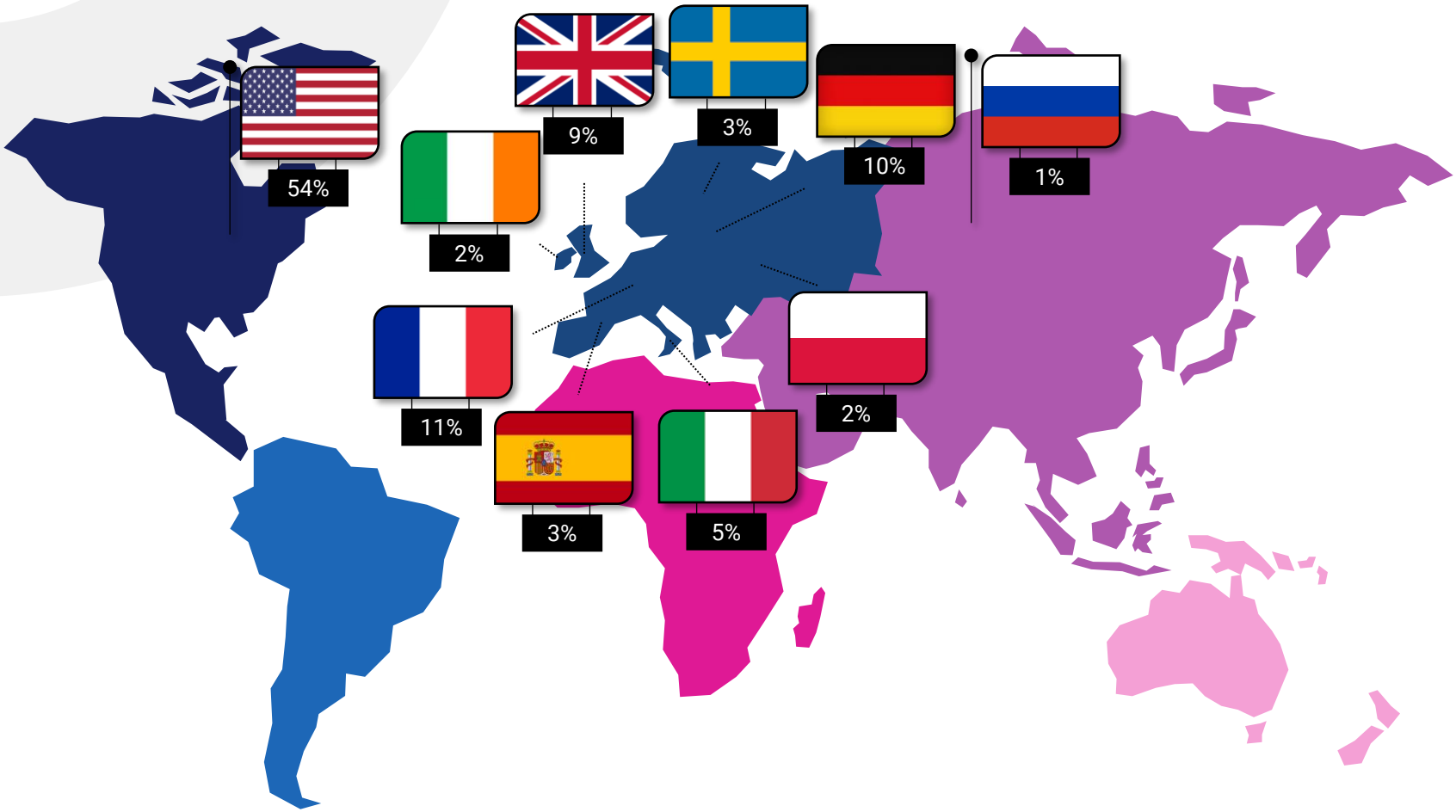

Figure 17 Country of Residence

Company Size, Job Functions, Seniority and Industry

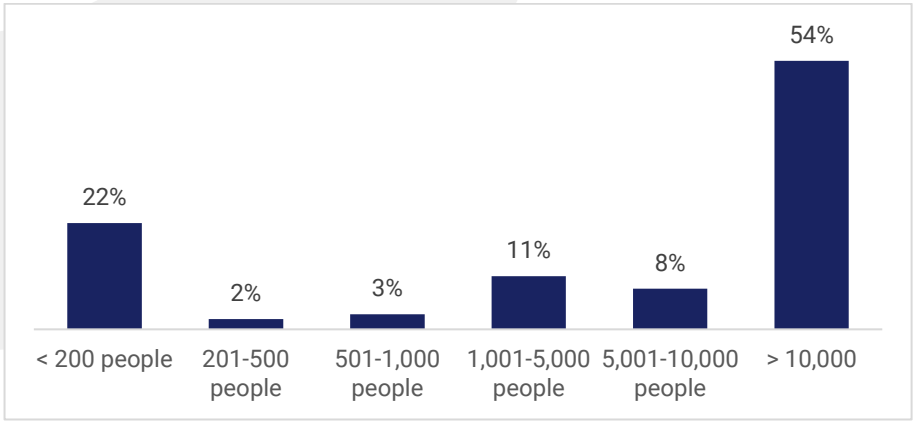

Figure 18 Company size

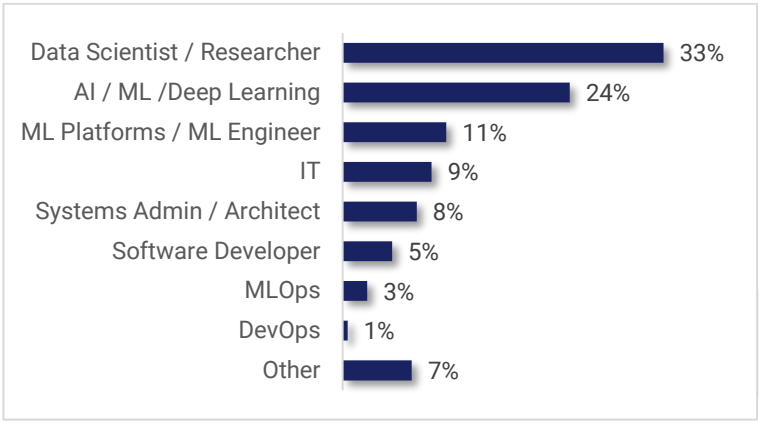

Figure 19 Job Function

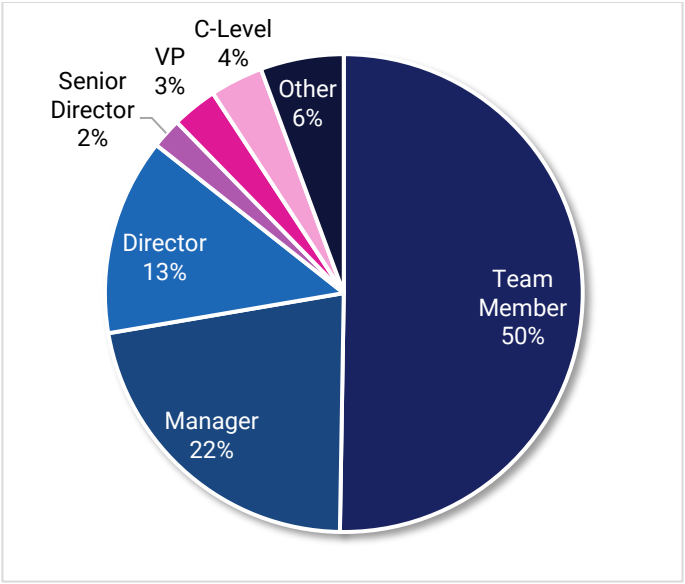

Figure 20 Job seniority

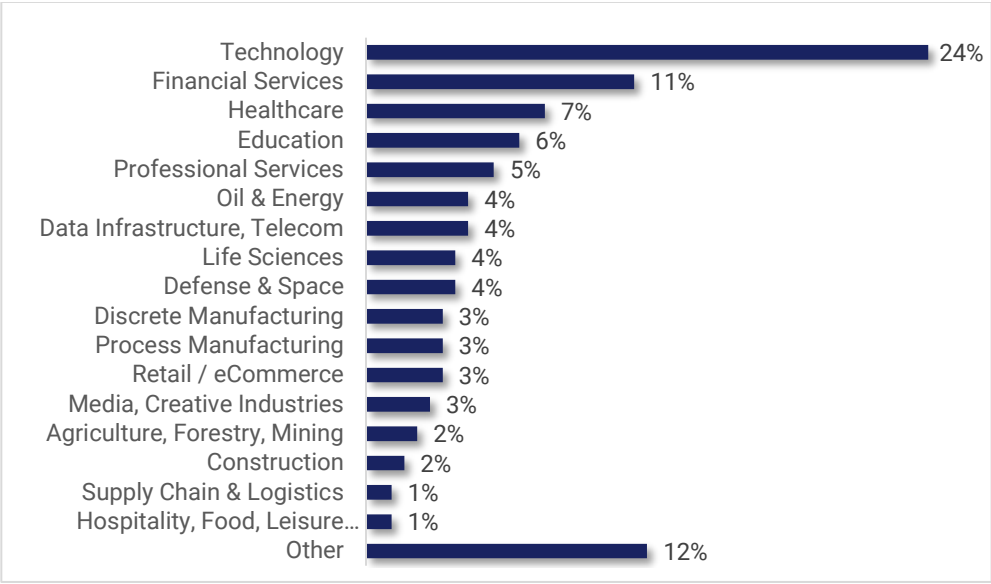

Figure 21 Industry

Actionable Steps Based on the Key Findings

These survey findings indicate that an organization's ability to move ML models to production on time and on budget is largely dependent on effective allocation and high utilization of GPU resources. A logical next step is to evaluate existing AI orchestration tools and processes to find areas of improvement.

If you are one of the x% who have infrastructure or compute challenges, Run:AI provides a cloud-native compute resource management platform for the AI era. With Run:AI, data scientists get access to all the pooled compute power they need to accelerate AI experimentation - whether on-premises or cloud. The company's Kubernetes-based platform provides IT and MLOps with real-time visibility and control over scheduling and dynamic provisioning of GPUs – and gains of more than 2X in utilization of existing infrastructure.

Enterprises seeking to compare their AI maturity with their peers and set actionable steps as their AI initiatives scale should consult the AI Infrastructure Maturity Model. The Model sets out six levels that encompass the milestones achieved on the path to a transformational level of ML maturity, particularly as it relates to mastering the many infrastructure challenges of productizing AI.

Request a Demo