FIDE PARTNERS

Impact of Generative AI on the Public and Private Cloud

Whitepaper

March 2025

Antonio Mosquera, Anna Krivonozhko, Juliana Nieto, Tomás Blasco

FIDE PARTNERS

Impact of Generative AI on the Public and Private Cloud

Tel: +44 2038 187213
info@fidepartners.com
www.fidepartners.com

March 2025

Whitepaper

Inside

Overview

The rapid growth of artificial intelligence (AI) is transforming cloud computing, driving innovation and redefining public, private, and hybrid environments. AI, with its demand for massive computational resources to train large-scale models, is pushing public clouds to address challenges like cost efficiency, data privacy, and latency. Private clouds are increasingly adopting specialised hardware, such as GPUs and TPUs, to manage the intensive requirements of training generative models while maintaining data control. Hybrid clouds, enabled by AI-driven automation, are emerging as a solution to balance scalability, cost, and security. This evolving synergy is reshaping cloud infrastructures and enabling the next era of AI-powered innovation

Cloud and AI: the increasing need for computational power

Introduction

The cloud market encompasses public, private, hybrid, and multi-cloud models, each designed to meet the diverse and evolving needs of modern enterprises and the public sector.

Public cloud services, delivered by leading providers such as AWS, Microsoft Azure, and Google Cloud, are popular for their cost-effectiveness (e.g., pay-as-you-go pricing models), primarily for low-capacity workloads, scalability to handle fluctuating workloads, and flexibility to support a wide range of applications. Dominating in key segments like SaaS, IaaS, and PaaS (software-, infrastructure- and platform-as-a-service), these platforms have become the backbone of modern enterprise IT, enabling organisations to innovate rapidly and efficiently.

Private cloud solutions, whether managed on-premise or by vendors like VMware or IBM, offer organizations greater control over their data, enhanced security measures, and extensive customisation options, making them more suitable for businesses with stringent regulatory or operational requirements.

Hybrid cloud combines elements of on-premise, private, and public cloud environments, enabling secure and efficient data and application transfers across these infrastructures. This model provides the security and control of private clouds with the scalability of public cloud resources.

Multi-cloud strategies involve combining services from multiple cloud providers within a single architecture. This approach enhances redundancy and improves flexibility by leveraging each provider's strengths. Often combined with hybrid cloud setups, multi-cloud strategies help organisations avoid vendor lock-in and create customised solutions for specific workloads.

Rise of Generative AI

Generative AI is transforming the cloud market by driving rapid growth in demand for computational power and adaptable, scalable infrastructure. Across the US and Europe, more than 54% of organisations identify AI as a core driver of their cloud strategies, with adoption rates exceeding 60% in industries such as banking and manufacturing (Wipro).

Looking ahead, global AI spending is expected to grow at a 29% CAGR from 2024 to 2028 and reach USD632bn by 2028 (IDC). This growth is expected to be mostly driven by generative AI with spending on related solutions growing at a 59% CAGR within the same period and reaching USD202bn by 2028 (IDC). Foundational AI models are expected to power 70% of NLP use cases by 2027, up from just 5% in 2022 (Gartner)

As a result, cloud computing faces increasing demand to support AI's intensive computational needs while maintaining scalability, compliance and data security.

Figure 1 – Comparison of cloud solutions

	Public	Private	Hybrid
Capex	○	◑	◔
Opex	◕	◔	◑
Security & compliance	◔	◕	◑
Management complexity	◔	◑	◕
Scalability	◕	◔	◑

Source: Fide Partners' analysis and research

Fuelling AI innovation: key resources and tools for AI deployments

The creation and deployment of AI models is a complex process that demands substantial hardware and software resources. These include computational capacity for tasks like training, inference, and fine-tuning, which are essential to developing and optimising AI models. Figure 2 (below) illustrates the interplay between resources used in AI model development.

Hardware viewpoint

AI workloads rely on various processors, such as CPUs, GPUs, and TPUs, each better suited to different stages of AI computing processes. GPUs, for example, are well suited to handle parallel processing demand, while TPUs have a better fit for deep learning thanks to optimised efficiency. Efficient data storage is another vital component, as AI training needs quick access to large datasets. Data lakes, warehouses, and other cloud storage solutions ensure scalability for fluctuating data volumes. High-speed networks facilitate rapid data movement between storage and processing nodes, which is especially critical in distributed AI systems.

Software viewpoint

The AI software stack includes frameworks like TensorFlow and PyTorch, which provide tools for model design, training, and deployment while leveraging specialised hardware to ensure efficiency and scalability in AI workflows.

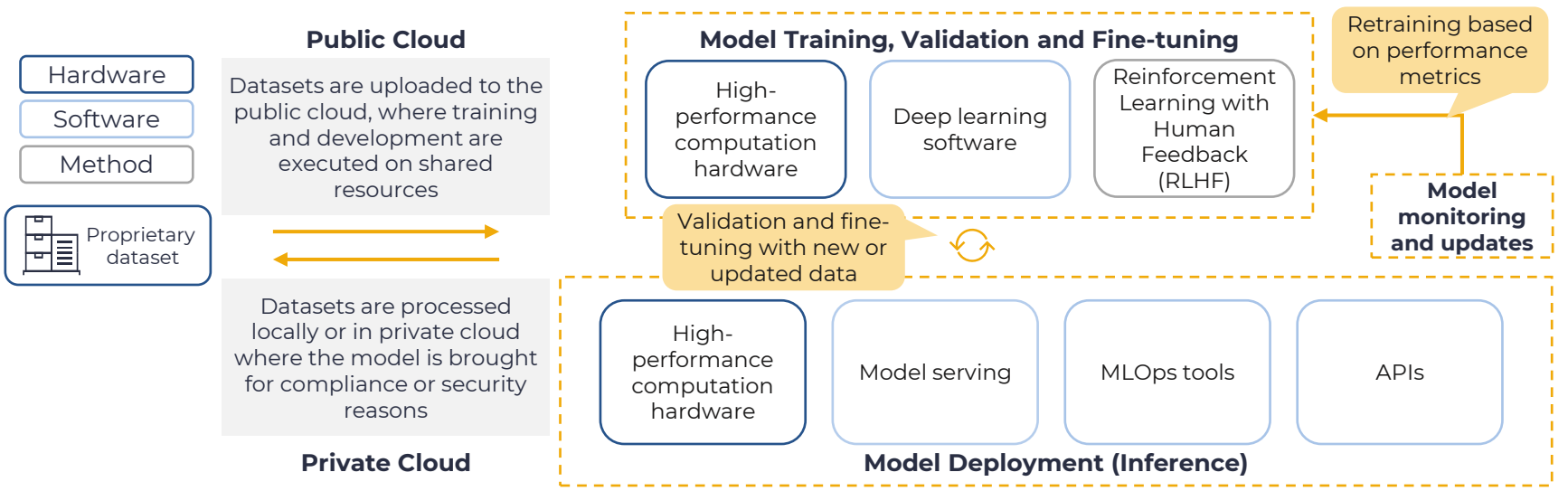
MLOps platforms like Azure ML and Vertex AI manage data collection, model training, versioning, and deployment, automating much of the process and enabling continuous model optimisation.

APIs provide access to pre-trained models, data, and computational resources, allowing developers to add AI features without starting from scratch. Initially proprietary and closed, these APIs made developers use specific cloud platforms, historically strengthening the reliance of developers on them. Over time, the APIs were gradually opened and made public, lowering the dependence of the clients on one single provider and favouring multi-cloud and hybrid models adoption.

The role of cloud service providers

Cloud service providers offer specialised hardware and software instances, enabling access to high-performance computational resources without requiring large upfront investments. While the value propositions of public, private, and hybrid cloud models differ, all of them appeal to those seeking to utilise AI capabilities without the need to build their own infrastructure or develop specialised in-house expertise, lowering entry barriers and accelerating AI adoption. Between 2018 and 2023, the global business AI adoption rate among organizations remained around 50%, substantially increasing to 72% in 2024 (Statista). This is largely driven by the increasing use of GenAI which jumped from 33% in 2022 to 65% in 2023 (McKinsey).

Figure 2 – Simplified diagram of resource usage for AI model development



Source: Fide Partners' analysis and research

The value proposition of public cloud for AI: accessibility, convenience and scalability

The public cloud has been transforming AI scalability for several years now, driven primarily by flexible APIs and dynamic resource allocation. Hyperscale APIs, offered by AWS, Google Cloud, and Microsoft Azure, simplify access to specialised AI functions, enabling businesses to quickly build and integrate advanced capabilities like natural language processing and computer vision. Over 50% of generative AI startups rely on these APIs to speed up development and reduce infrastructure complexity (McKinsey). APIs are fundamental to the public cloud, as they enable seamless integration, interoperability, and functionality across various services. Hyperscalers play a pivotal role in the rapid deployment and dissemination of these APIs, ensuring businesses can efficiently leverage cloud-based innovations. However, the speed at which hyperscalers release APIs and extend their availability beyond their own cloud ecosystems varies significantly, directly impacting the benefits of the public cloud, such as interoperability, scalability, and service availability across multiple platforms.

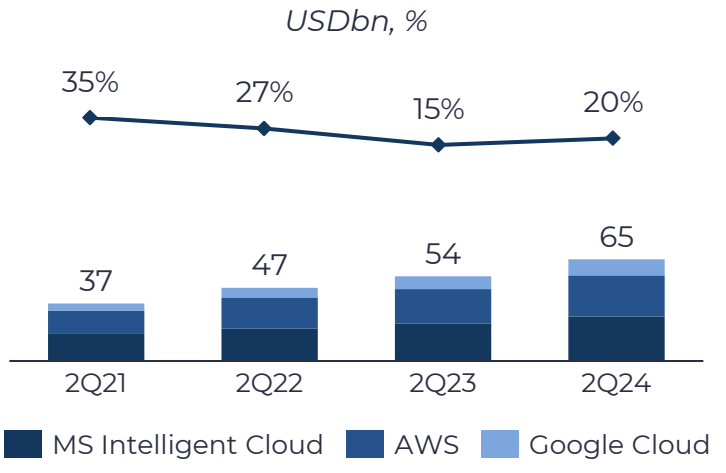
Another value proposition of the public cloud model lies in its cost-effectiveness, primarily for smaller-scale workloads.. The public cloud providers offer a pay-as-you-go model, which allows companies to scale resources based on immediate needs without large, upfront investments in hardware (for the on-premise solutions) or planning for the exact capacity to be contracted with a private cloud provider. This model benefits startups and enterprises, with unpredictable workloads or short development lifecycles, while customers with more stable demand may find alternative models more cost-efficient, especially in the long-term.

Additionally, public cloud platforms often provide access to cutting-edge processing tools essential for scaling AI models, driven by innovations from providers like Nvidia and Microsoft. For example, NVIDIA's DGX Cloud on AWS allows enterprises to utilise the latest Blackwell GPUs for AI training and inference. These GPUs offer unprecedented processing power, reaching up to 414 exaflops, which is 375 times faster than the current top supercomputers, supporting data-heavy tasks like generative AI.

Hyperscalers leverage a vast global infrastructure and large-scale data centres to handle massive workloads and deliver AI services more efficiently than alternative smaller providers. Their ability to purchase hardware in bulk, automate resource management, and optimise infrastructure allows them to offer a wide range of solutions tailored to different customer requirements. Meta, Microsoft, Google and AWS jointly purchased around 400,000 NVIDIA H100 GPUs[1], accounting for 66% of global shipments in 2023. In addition, these solutions are supported by streamlined business processes and are often managed by the provider, ensuring high convenience for customers.

The benefits of using public cloud for AI deployments have driven its adoption, which has been evident in the revenue growth of public cloud providers as it began to recover after slowing down between 2022 and 2023. In October 2024, Microsoft announced that its AI business was on track to surpass an annual run-rate of USD10bn by the end of the year and that its Azure and other cloud services revenue grew by 33% YoY, of which 12 p.p. were contributed by AI services. Similarly, Google Cloud's revenue in the third quarter of 2024 surged by 35% YoY driven by AI, and AWS's 19% YoY revenue growth for the same period was fuelled by its GenAI services and ML products, growing three times faster than AWS itself.

Figure 3: Q2 cloud revenues and YoY growth of the main public cloud providers (2021-2024)



Source: Annual reports 2021-2024

Unlocking the potential of the private cloud: empowering hybrid solutions for AI

The private cloud model offers a distinct value proposition for developers and organisations. It provides a dedicated infrastructure with enhanced security and privacy, allowing for full control over resources and predictable costs.

It benefits industries like finance, healthcare, manufacturing and the public sector, primarily due to its enhanced data security, compliance capabilities and control over sensitive information. In finance and healthcare, private cloud environments ensure compliance with regulations, safeguarding critical data. In manufacturing, it securely centralises IoT and predictive maintenance data, protecting intellectual property and allowing for real-time, secure data analysis. For government organisations, the private cloud supports data sovereignty, helping control access to critical information.

At the same time, private cloud environments present certain challenges for AI workloads. An on-premise private cloud environment involves high upfront costs (capex), required for dedicated infrastructure such as processing units, as opposed to the opex-driven public cloud model. Rapid scalability is often constrained in the private cloud model, where resource expansion relies on available physical infrastructure, making it less flexible for AI's resource-intensive and fluctuating demands.

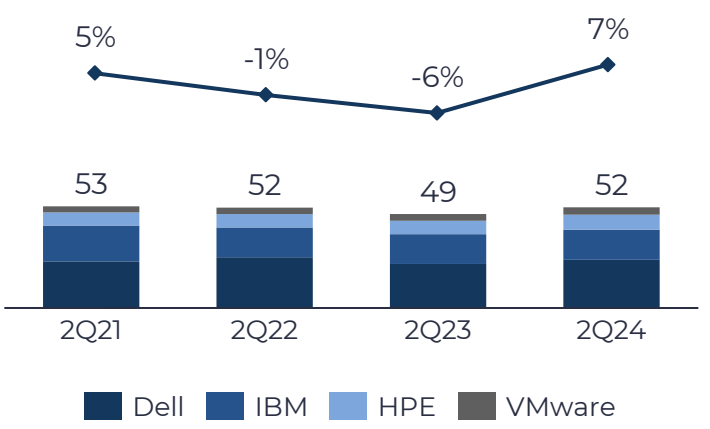
On the other hand, a private cloud environment hosted by a third party like IBM, VMware or HPE, while also providing the customer with dedicated resources and customised solutions, does not involve high upfront costs compared to the fully on-premise model. Its cost structure is more predictive compared to the public cloud model, which can be an optimal solution for organisations with predictable and less fluctuating workloads. However, while it is possible to scale resources up or down, in the private cloud model this would require physical hardware changes that can be expensive and time-consuming. This is where the hybrid cloud model steps in, leveraging the advantages of public and private cloud solutions.

The hybrid cloud setup allows sensitive data to remain in a private environment while utilising public cloud for intensive, scalable AI tasks. In a typical hybrid model, proprietary and sensitive data is stored and processed on-premise or in the hosted private cloud. At the same time, AI models are trained and tested in the public cloud using large, non-sensitive datasets, leveraging access to the latest APIs and scalable compute resources when required. The combined use of private and public clouds has emerged as a preferred option for many organisations, with 78% of enterprise IT decision-makers utilising hybrid cloud for building and deploying generative AI solutions (Enterprise Strategy Group).

This trend has been reflected in the revenue of private and hybrid cloud providers. IBM delivered a strong performance in the second quarter of 2024, directly attributed to its AI-related services, with its bookings for generative AI-related services exceeding USD2bn. In the third quarter of 2024, HPE saw strong growth for its AI server and AI system segments, with the AI server segment achieving a record of USD1.5bn in revenue and contributing 16 p.p. to the 32% YoY growth of the overall service segment. Similarly, Dell's infrastructure segment reached USD11.4bn in revenue in the same period, marking 34% YoY growth, with AI server sales playing a critical role.

Figure 4: Q2 revenues and YoY growth of the main private cloud providers (2021-2024)

USDbn, %



Source: Annual reports 2021-2024

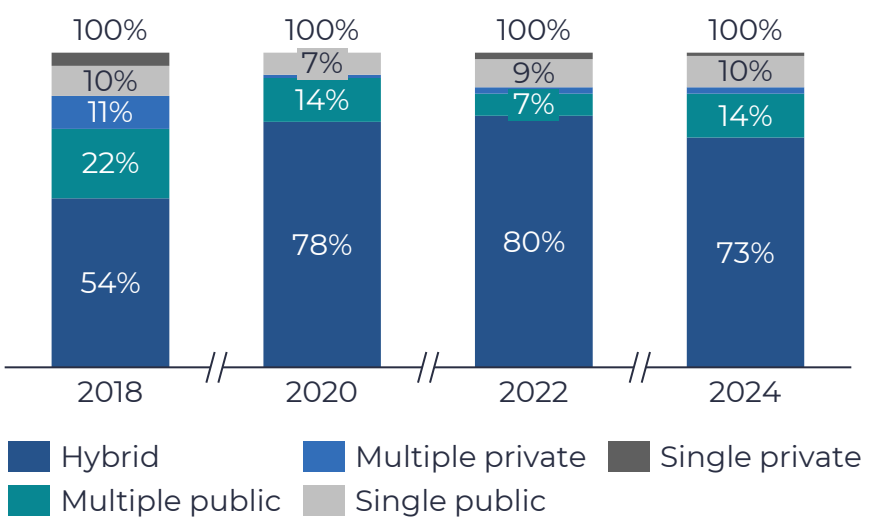
Moving forward: the rise of hybrid and multi-cloud models

The AI market is expected to keep growing driven by the increasing adoption of generative AI and ML models across industries. As the market continues to grow, businesses will need to remain agile to grasp the benefits of generative AI, which involves considering the evolving and demanding hardware and software requirements to deploy future-proof AI solutions. Developers increasingly pay particular attention to long-term sustainability, focusing on flexibility and scalability yet prioritising long-term cost optimisation, maintaining control over sensitive data and complying with security and privacy issues according to the latest regulations.

The public cloud model has undeniable advantages, allowing for easy scalability and a wide range of APIs, which will likely remain the preferred choice for early-stage developers and customers facing highly fluctuating workloads. At the same time, the advantages of the private cloud have been gaining momentum, among other factors, due to the actively developing regulatory context shaping the AI sector. These regulations will become only more extensive, definite and stringent as the AI adoption increases.

Another relevant consideration for developers moving forward will be cost-effectiveness and access to GPUs. The general convenience and accessibility of the public cloud model do not always imply the optimal choice from the cost perspective in the long term, especially once the workloads become more predictable, and organisations can benefit from a more predictable pricing structure of private cloud solutions. Hybrid cloud is therefore expected to continue strengthening its position as a solution enabling compliance, higher protection of sensitive data at a required scale, and predictable costs. Moving forward, hybrid cloud is expected to dominate the cloud market with 90% of the organisations expected to be using hybrid cloud by 2027 (Gartner). In addition, the multi-cloud model is also expected to see increasing demand as developers try to avoid vendor lock-in and find an optimal combination of solutions, tailored to their needs.

Figure 5: Cloud infrastructure adoption (2018-2024)



Source: Flexera

Regulatory considerations

Regulations impacting AI cloud computing mainly focus on data privacy laws (e.g., GDPR, CCPA), which restrict how data is managed in AI systems and cloud environments, especially with massive cross-border datasets. Emerging frameworks, like the EU AI Act, address AI ethics, including transparency and bias, which could impact cloud-based AI services. This will likely impact sensitive industries' choice of private and hybrid cloud solutions for greater data control. Public cloud providers, in turn, are adapting to meet data sovereignty needs through localised processing and compliance solutions. To meet regulatory requirements, businesses can adopt a multi-cloud strategy that leverages different providers for specific AI services by placing workloads in different environments. This approach facilitates compliance with evolving regulations.

AI demand is shaping the cloud market through evolving computational needs and increasing adoption rates. The race to implement AI and reap its benefits is equally growing among startups, enterprises and public sector. Public cloud model is expected to continue being a preferred choice for developments with short lifecycles or unpredictable workloads. At the same time, hybrid cloud and multi-cloud strategies are set to gain traction, leveraging unique competitive advantages of different models and providers, and ensuring scalability, compliance, data security and cost effectiveness.

Relevant case studies

Tesla: Private cloud in the automotive industry

Headquarters

Tesla operates a sophisticated private cloud infrastructure to power its AI-driven applications. The Cortex AI supercluster is an **on-premise private cloud** with ~70,000 AI servers, which ensure **data privacy, reduces reliance on external providers**, and **optimises cost and performance** for Tesla's proprietary AI tools

Tesla integrates **generative AI** into its private cloud infrastructure to optimise **autonomous driving, vehicle maintenance, design, and energy solutions**:

- **Autonomous driving:** Tesla uses generative AI to simulate complex driving environments, enabling safer and more efficient training for its self-driving systems
- **Predictive maintenance:** AI analyses vehicle data to predict potential issues, reducing downtime and improving reliability
- **Main hardware provider:** NVIDIA

Tesla's private cloud model ensures a full control of its proprietary data while complying with global regulation

Netflix: Public cloud in the VoD (Video on Demand) industry

Headquarters

Netflix operates exclusively in a **public cloud environment** to power its streaming platform, big data, analytics, and recommendation systems. Netflix hosts and delivers content through AWS public cloud targeting **seamless playback for millions of users worldwide**

Netflix integrates advanced cloud, **AI and ML technologies** into its operations to ensure user engagement and retention through **personalised recommendations**:

- **Data lake:** Netflix uses a cloud-based data lake to store data such as streaming logs, interaction data and metadata
- **AI algorithms:** Netflix employs Machine Learning models hosted on AWS to analyse viewing patterns and generate personalised content suggestions for each user
- **Main cloud provider:** AWS

Netflix's public cloud approach enables scalability, global reach, and flexibility to handle fluctuating demand

JP Morgan: Hybrid cloud in the Finance and Banking industry

Headquarters

JP Morgan Chase has adopted **a hybrid cloud approach.** This strategy **mitigates risks of vendor lock-in and provides greater control over sensitive operations**. Compared to on-premises solutions, hybrid clouds deliver superior scalability and innovation capabilities without the high costs

JPMorgan integrates **AI for improved risk management, customer insights,** and **operational efficiency**:

- **Risk management:** JP Morgan uses ML models, NLP, real-time monitoring and behavioural analytics for fraud detection
- **Customer insights:** AI-driven analytics help personalise financial products and services for tailored recommendations
- **Operational efficiency:** JP Morgan deployed an internal AI assistant using AWS ML and GenAI platforms
- **Main cloud provider:** IBM Cloud

JP Morgan's hybrid cloud supports its bet on continuous AI innovation and the control of the sensible data it handles

Source: Fide Partners analysis and research

FIDE PARTNERS

Contact:

Antonio Mosquera

Mail:
antonio.mosquera@fidepartners.com

Anna Krivonozhko

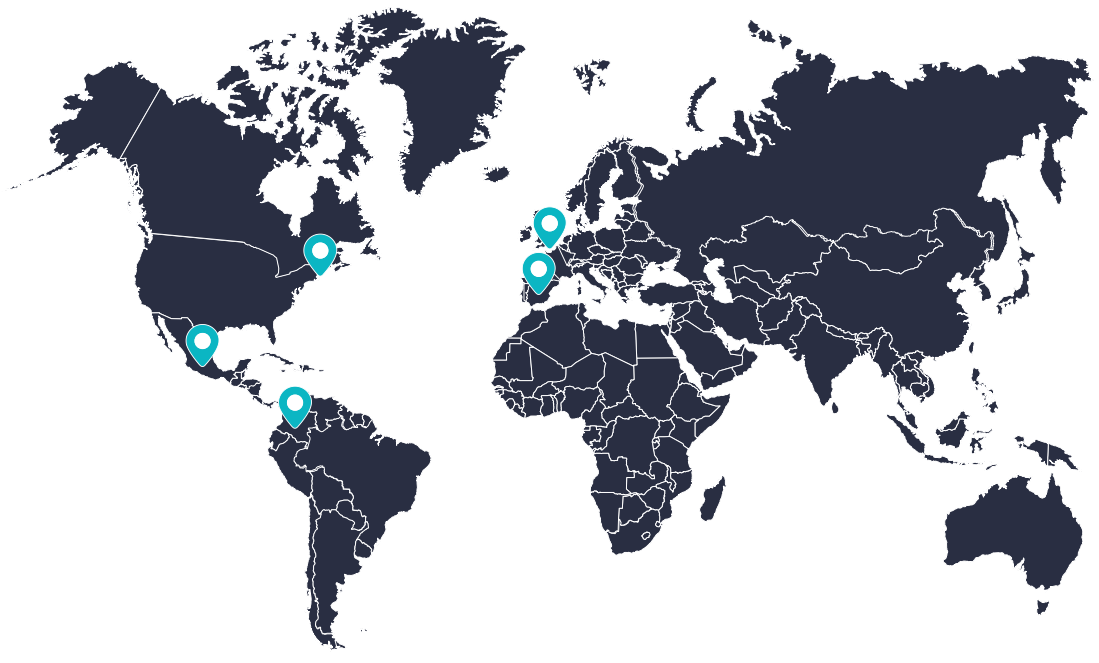
Mail:
anna.krivonozhko@fidepartners.com

Juliana Nieto

Mail:
juliana.nieto@fidepartners.com

Tomás Blasco

Mail:
tomas.blasco@fidepartners.com



Fide Partners offers specialised expertise to investors seeking to leverage the rapid rise of AI adoption, a key driver of growth in the data centre industry. As AI continues to reshape the data centre market and influence demand from cloud service providers and colocation platforms, our insights help you navigate this evolving landscape. Visit our website to learn more: **www.fidepartners.com**

Direction:

London:
Aviation House
125 Kingsway, 6th Floor, Office 107
WC2B 6NH, London
United Kingdom

Madrid:
C/Don Ramón de la Cruz, 6, 1°
28001 - Madrid
Spain

Bogotá:
Carrera 11A #98-50
Ofc. 704, Edificio Punto99
110221, Bogota
Colombia

Boston:
50 Milk Street,
Planta 15, C.P. 02109.
Boston, MA