IRON MOUNTAIN®
DATA CENTERS

The twin infrastructure impacts of generative AI and how to deal with them

DIGITAL REPORT 2023

IN ASSOCIATION WITH:

HKS

IRON MOUNTAIN®
DATA CENTERS

THE TWIN INFR IMPACTS OF G

AND HOW TO DE

PRODUCED BY: **IRON MO**

RASTRUCTURE
ENERATIVE AI

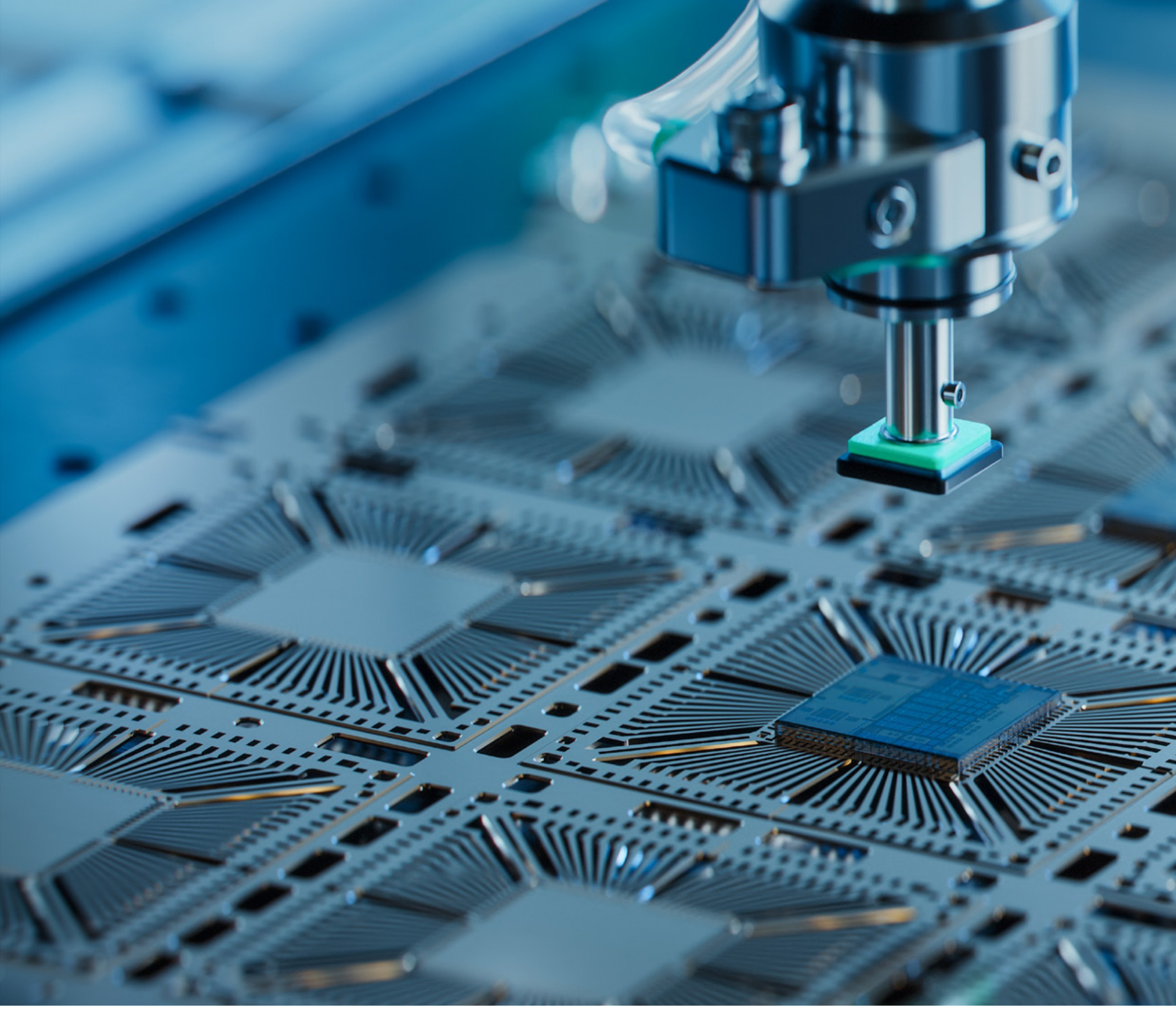
EAL WITH THEM

OUNTAIN DATA CENTERS

Sustainability leader *Iron Mountain Data Centers* is investing heavily in next generation carbon reduction measures which will be vital to sustain the exponential growth of generative AI

2023 will be looked back on as the year generative AI burst into the public consciousness and businesses frantically adapted their models to accommodate it. The ability of Large Language Models (LLMs) to bridge the linguistic gap between humans and machines has caught the popular imagination and raised awareness of the potential to automate and improve many aspects of our lives.

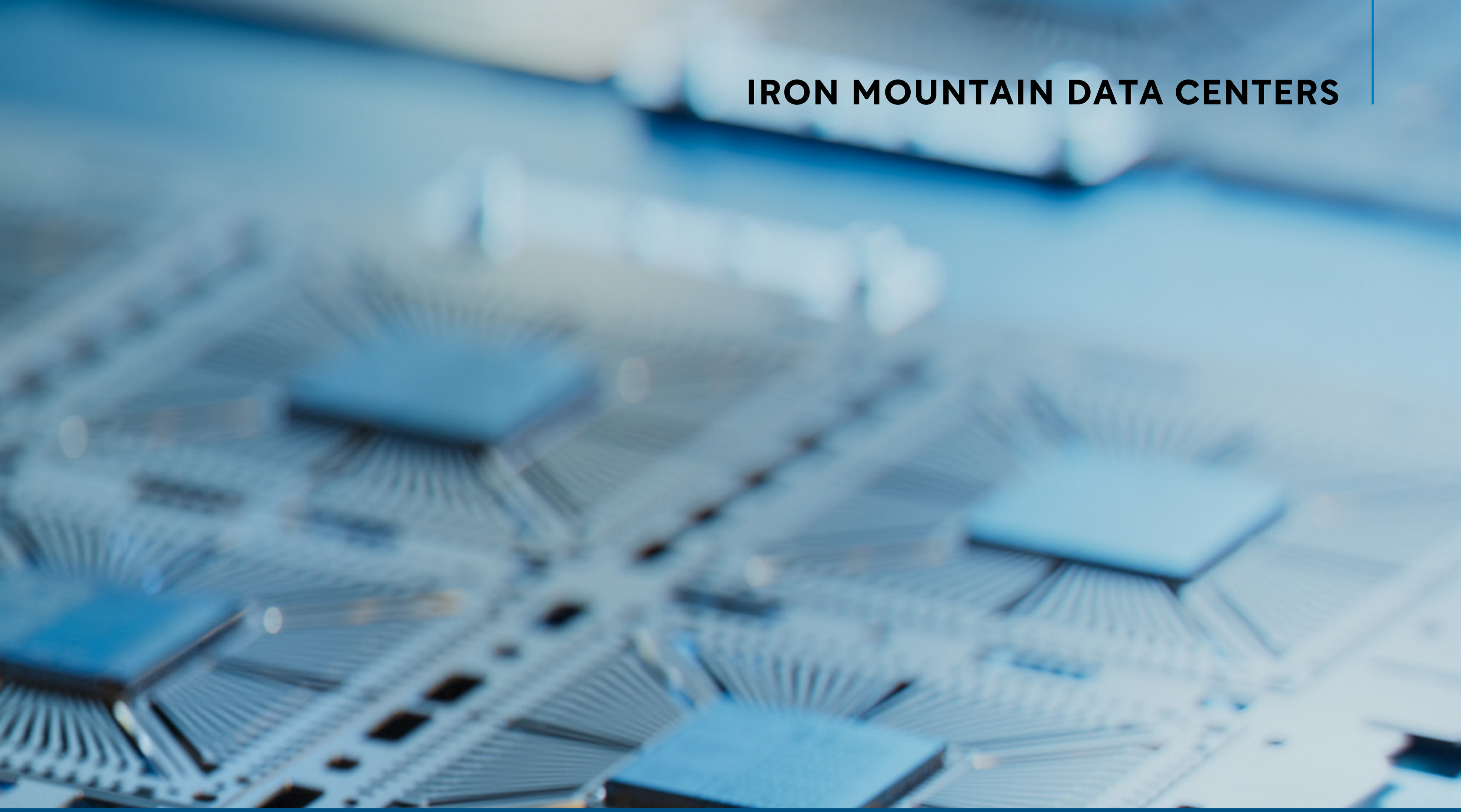
In this initial flurry of speculation it can be difficult to find reliable forecasting models on which to base sound business decisions. However, in data centre infrastructure the impacts are more predictable than most, and Iron Mountain Data Centers (IMDC) believes they will drive an industry-wide design revolution fuelled by sustainability. Smart data centre users undertaking AI investment should be aware of this and start planning for it now.



Power surge

It is clear to Iron Mountain Data Centers that by far the greatest challenge in supporting generative AI is the huge surge in power loads. Generative AI models use graphics processing unit (GPU) chips which require 10 to 15 times the energy of a traditional CPU. Many models have billions of parameters and require fast and efficient data pipelines in their training phase, which can take months to complete. ChatGPT 3.5, for instance, has 175 billion parameters and was trained on more than 500 billion words of text. To train a ChatGPT 3.5 model requires 300 to 500MW of power. Currently, a typical data centre requires 30 to 50MW of power. One of IMDC's larger campuses, in Northern Virginia, has capacity for 10 data centres on it. The whole of the power load for this campus would be required to train ChatGPT 3.5. While LLMs are definitely at the most power-hungry end of the generative AI boom, every generative model IMDC has worked with has processor and power needs which grow exponentially, either doubling or tripling each year.

Forecasting the power requirements of generative AI over time is hard to do with any accuracy, but most analysts agree that it will ramp up current requirements hugely. If one estimates current data centre compound growth at a relatively modest



AI'S APPETITE IN ACTION

IMDC provides the infrastructure for many High-Power Compute (HPC) configurations running generative AI, and has developed specialist facilities that meet their needs. High-density power, modular architecture, high-bandwidth training – input – and inference – output – connectivity and advanced cooling are all critical factors for customers.

Healthcare
One of IMDC's healthcare customers has developed a supercomputer for AI-driven imaging apps. While the total consumption is not massive and training cycles are much shorter than for LLMs, the data growth curve of this supercomputer since it was first built in 2018 has been steep. It began with just 10,000-50,000 images and achieved 85% accuracy. Now it uses up to half a billion images with accuracy of 95% and runs 50,000 deep learning training experiments per month. Despite the compactness and efficiency of the GPUs, a few racks in the data centre have become a full module of 60 racks with 26 petabytes of processing power storing close to two billion datasets. In less than two years a petaflop of processing power will be needed.

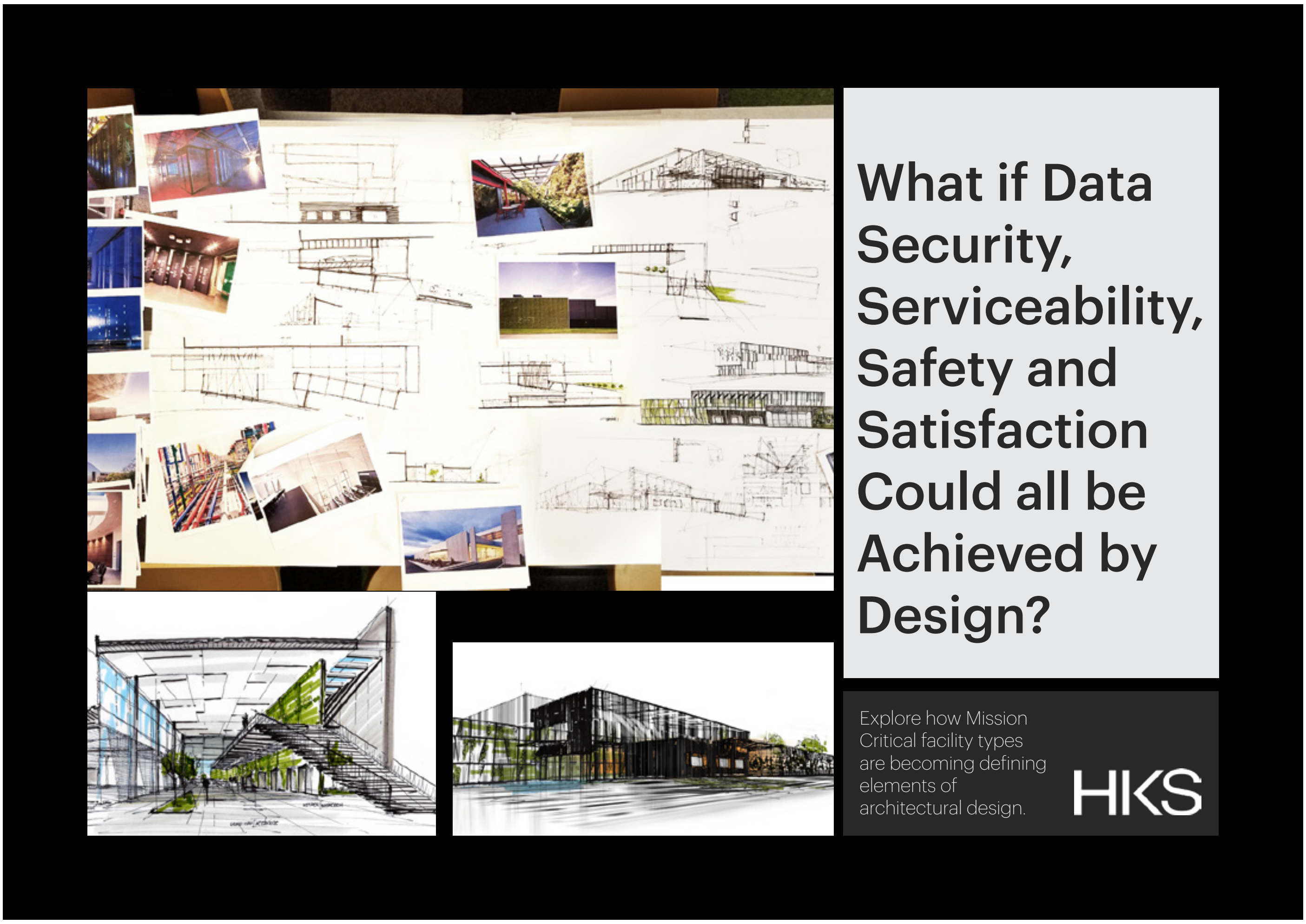
Research
The Computational Research Accelerator department at Arizona State University was running out of network ports, space and power for 'Agave', its supercomputer, so they built a new supercomputer called 'Sol' in 2022 in one of IMDC's Phoenix data centres. Sol is a Dell-built system spanning 178 nodes. It uses AMD Epyc 7713 CPUs, consisting of around 18,000 cores, with the bulk of the nodes carrying 512GB of memory and five large-memory nodes equipped with 2TB. It has 56 GPU nodes with quadruple Nvidia A100, 80GB, GPUs each and four nodes with triple Nvidia A30, 24GB, GPUs. The system is networked with Nvidia's 200GB/s HDR InfiniBand and supported by four Petabytes of Dell BeeGFS scratch storage. The R&D potential of Sol is extremely exciting, and a steep physical growth curve is anticipated.

15%, global capacity will double in five years and quadruple in 10. With generative AI in the mix, CAGR could rise as high as 25%, tripling capacity in five years and increasing it up to tenfold in a decade. That is more than double the current growth rate. Enterprises, AI startups and Cloud Service Providers are already racing to secure data centre capacity for their workloads, with the hyperscale clouds leading the pack. This is happening fast. Analyst TD Cowen reported "a tsunami of AI demand" with 2.1GW of data centre leases signed in the US, a fifth of current total supply, in Q2 of 2023.

A mountain of e-waste

The second AI-generated challenge is at the back end; a stream of used equipment. AI is driving faster server innovation, particularly in chip design, and the latest AI chips such as the Nvidia H100 have had so many billions advanced against their manufacture and are in such short supply that they are even being used as debt collateral and made available for rent. While this refresh rate will be key to improving efficiency it will also – in tandem with the rise in capacity – increase the scale of e-waste.

E-waste is one of the fastest-growing waste streams in the world. By 2030, annual e-waste production is on track to reach a staggering 75 million metric tonnes. Global e-waste is thought to hold roughly US$60 billion-worth of raw materials such as gold,



> “To address the twin challenges of capacity ***growth and e-waste*** the industry will have to be at the top of its game”

palladium, silver and copper. However, just 17% of global e-waste is documented to be collected and properly recycled each year.

Looming climate targets

Combine these factors with the broader issues society now faces. These challenges will need to be addressed as the climate crisis deepens and zero emission targets loom. There will be unprecedented pressure on power grids to provide new electrical power for industries that are weaning themselves off fossil fuels. Iron Mountain Data Centers is a firm believer that generative AI in particular will be under intense environmental, and therefore

> “AI may be ***a key to solving the problem,*** not just for our own industry but for other sectors”

popular, scrutiny. To address the twin challenges of capacity growth and e-waste the industry will have to be at the top of its game.

Addressing the challenge: Carbon elimination and circularity

How should the industry react? As ever, by solving the problems one by one.

Low-to-no-carbon power sources will be the key to addressing power challenges. The power demands of generative AI will accelerate this focus and drive new innovations in microgrids and backup power sources such as battery, hydrogen and nuclear. Renewables will also be key. Most hyperscalers and a growing number of colocation providers have been growing the green grid and eliminating carbon to the point that today, hyperscalers are the biggest buyers of renewables in the world. On the colocation side, the Iron Mountain Group is now one of the top 20 renewable buyers in the world.

Data centre owners will now need to follow the leaders and we are already making headway. Following Google’s lead, two years ago IMDC committed to provide not just 100% renewables but 24/7 carbon-free energy – you can see how IMDC and its partners have gone about this in a recent documentary ‘Transforming our Future’. This is a major step up from ‘attributing’ power used to renewables credits, and we believe that this approach will in time replace the current year-by-year Virtual Power Purchase Agreement model.

When it comes to circularity, new chips and superfast GPUs will drive the

AI revolution, but what will happen to the old ones? For both efficient performance and impact reduction, Iron Mountain Data Centers says AI providers will need to check that IT asset lifecycle optimisation and recycling, remarketing and secure disposal are available. The industry has been fairly slow to integrate this, but this will accelerate, and IMDC is changing the shape of its business to be in a position to address this issue.

The IMDC Asset Lifecycle Management (ALM) division, which now covers 32 countries, sanitises more than three million drives a year and has generated in excess of US$1bn for clients via remarketing and recycling. Most recently, Iron Mountain invested a further US$200 million in acquiring Regency Technologies, which will add even more robust remarketing and recycling capabilities to support circularity for the world's largest digital businesses. Iron Mountain sees huge potential for this segment to service AI customers over the coming years.

The AI opportunity for the industry

In the same way that generative AI will revolutionise the industries that run its applications, it is set to revolutionise

STATS:

- To train a ChatGPT 3.5 model requires ***300-500MW*** of power
- With generative AI in the mix, data centre CAGR could rise as high as ***25%,*** tripling capacity in ***five years*** and increasing it up to ***tenfold*** in a decade
- By 2030, annual e-waste production is on track to reach a staggering ***75 million metric tonnes***
- The IMDC Asset Lifecycle Management (ALM) division, which now covers ***32 countries,*** sanitises more than three million drives a year and has generated in excess of ***US$1bn*** for clients via remarketing and recycling
- Click here to read the IMDC Infrastructure Service Sheet on the ***Top 10 considerations*** when planning AI infrastructure

the infrastructure industry that supports it. It promises to deliver immense economic value over the coming decade, but will also consume immense amounts of power. Many generative AI applications can be hosted in a specialised shared facility. Different models have different infrastructure requirements, but all share the need for high-density power, advanced cooling and modular design.

The scale of the power challenge does not mean it cannot be overcome. In an era in which Big Tech has displaced many oil giants in the list of the world's largest companies, innovation has accelerated and serious fact-based commitments have been made by industry leaders to tackle the climate crisis. In fact AI may be a key to solving the problem, not just for our own industry but for other sectors.

Data centre customers interested in developing generative AI applications should plan and invest early to keep ahead of the steep upward curve in power and space uptake. They should also pay close attention to new infrastructure design impacts,

> “Iron Mountain is now one of the ***top 20 renewable buyers*** in the world”

efficiency, energy sourcing and e-waste. This means scrutinising the energy track record and targets of their cloud or data centre provider and sharing data on climate target progress and day-to-day access to – preferably 24/7 carbon-free – renewables.

You can find more detail on the market forecasts and detailed infrastructure impacts of generative AI in IMDC’s Generative AI Solution Guide.

IRON MOUNTAIN®
DATA CENTERS

Iron Mountain Data Centers
85 New Hampshire Avenue, Suite
150, Portsmouth, NH 03801
03801

T 800-899-4766
www.ironmountain.com/
data-centers

POWERED BY:

BizClik