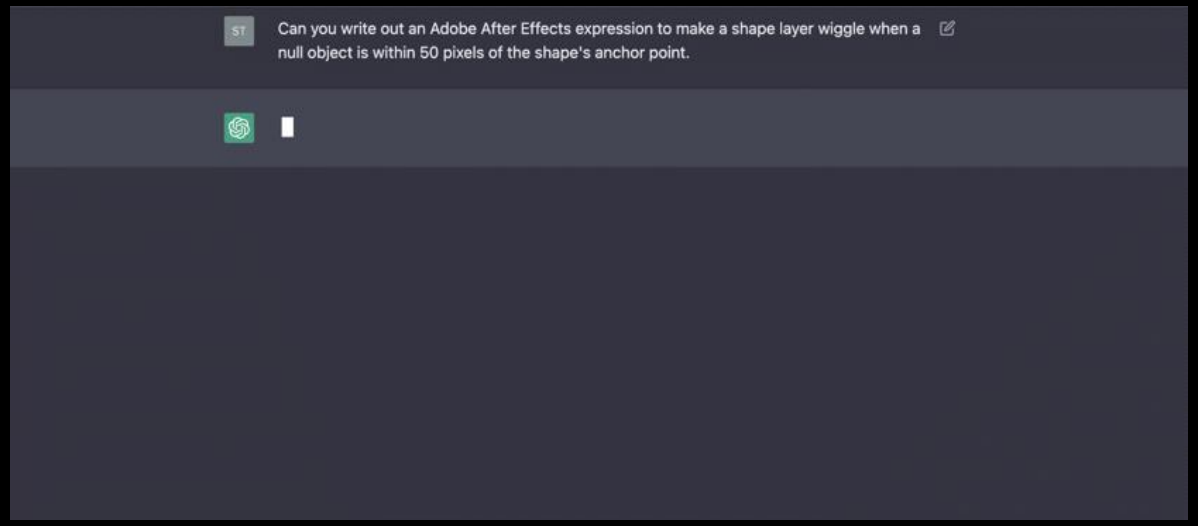
Accelerate your Workflows and Gain Competitive Advantage with AI Workstations

Ruchi Bhatia
Product Marketing Manager, Data Science & AI, HP

Allen Bourgoyne
Director, Product Marketing – Professional Visualization Solutions, NVIDIA

You've Heard About AI



CHATGPT

MIDJOURNEY

Amazon is Going 'Super Aggressive' on Generative AI

'Microsoft is back.' How AI Put the Five-Decade-Old Tech Giant on Top Again

Meta Debuts Code Llama 70B: A Powerful Code Generation AI Model

Google Says New AI Model Gemini Outperforms ChatGPT in Most Tests

Apple Explores A.I. Deals With News Publishers

The Components of AI Systems

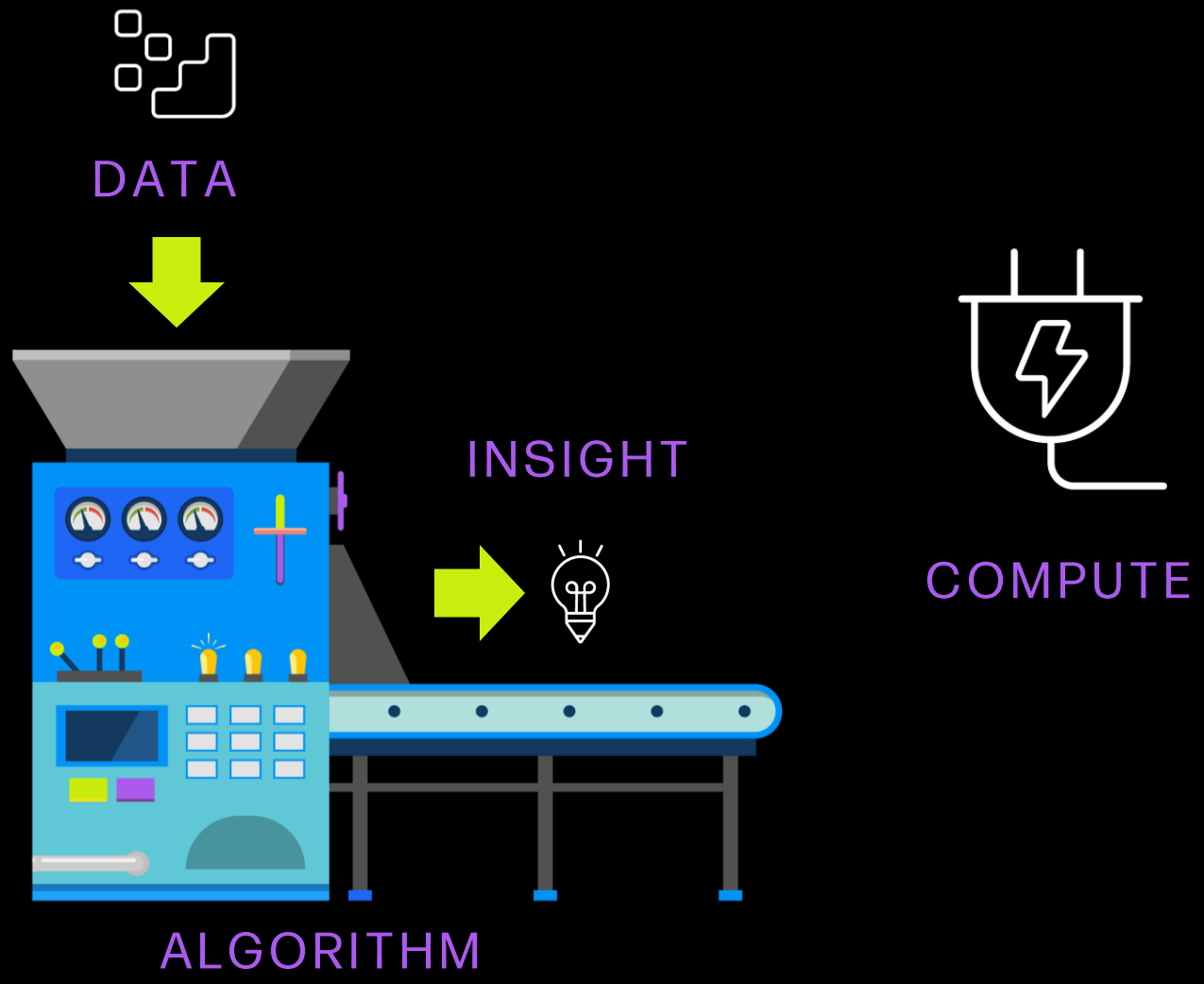
Putting together an AI system that generates insights

ALGORITHM

DATA

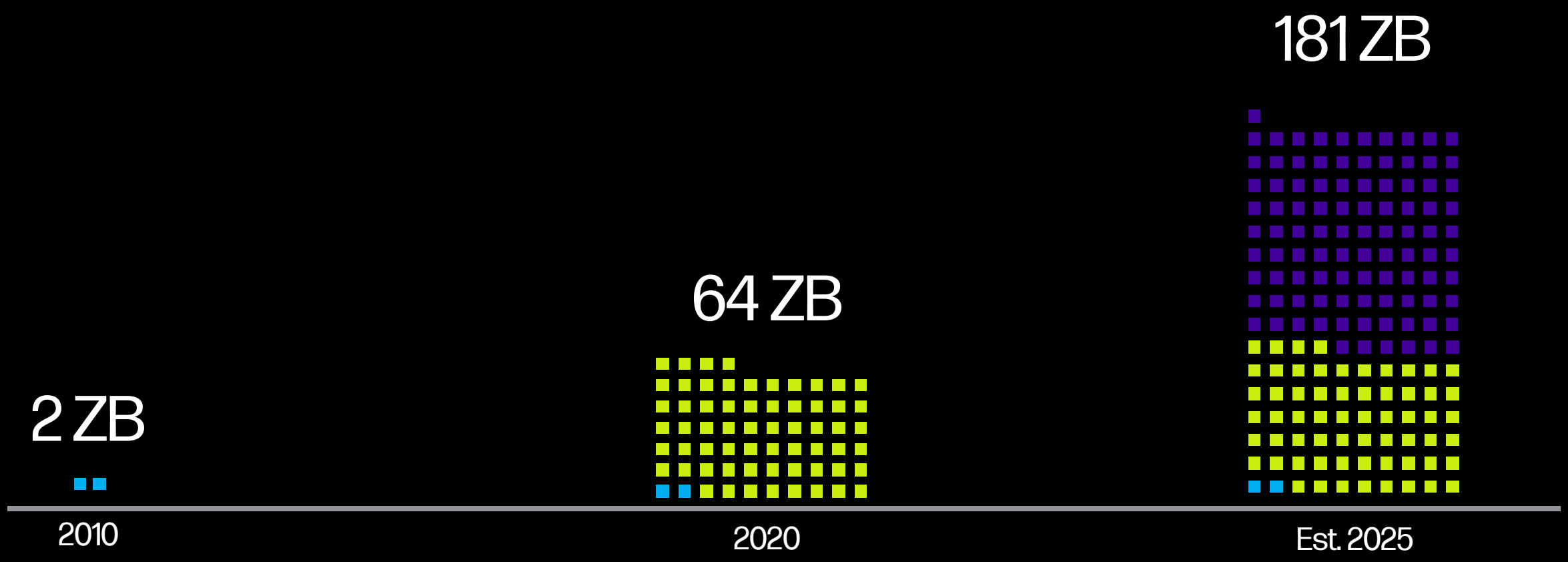
COMPUTE

AI: The Factory of Insights



Explosion in Data Generation

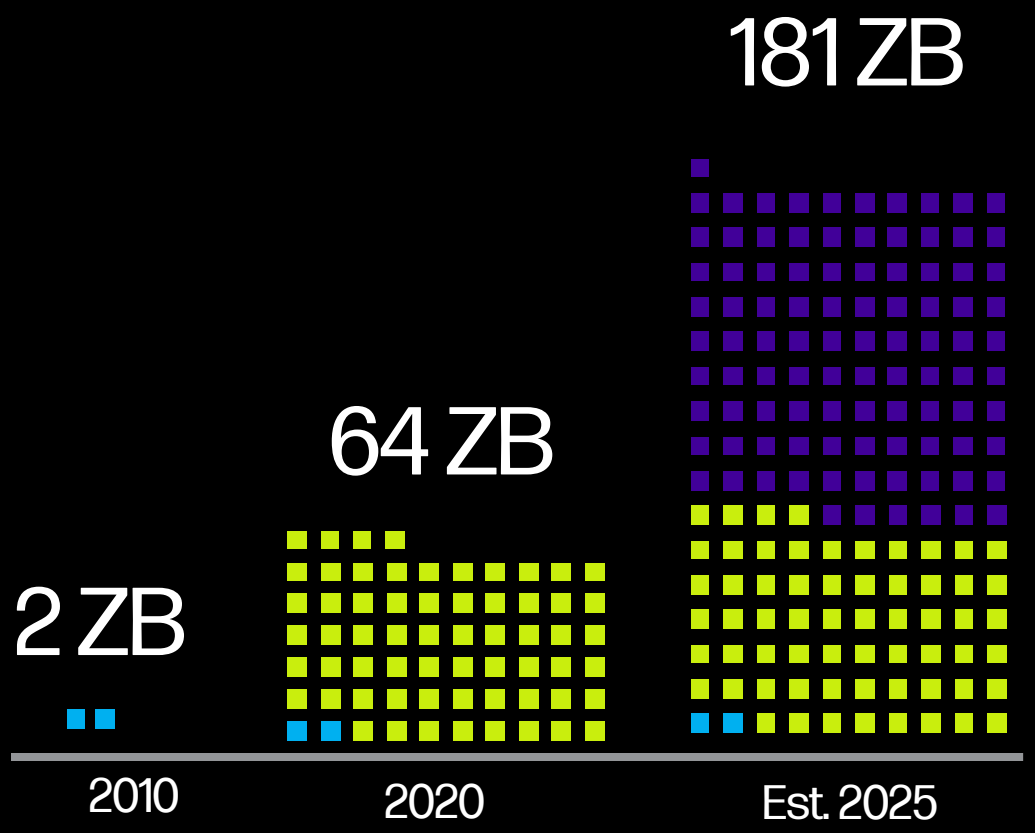
A trend that presents both challenge and opportunity



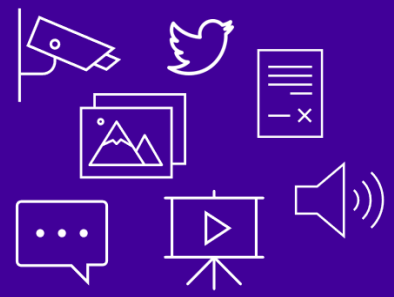
Total Annual Data created and stored (ZB[1])

Explosion in Data Generation

A trend that presents both challenge and opportunity



Unstructured



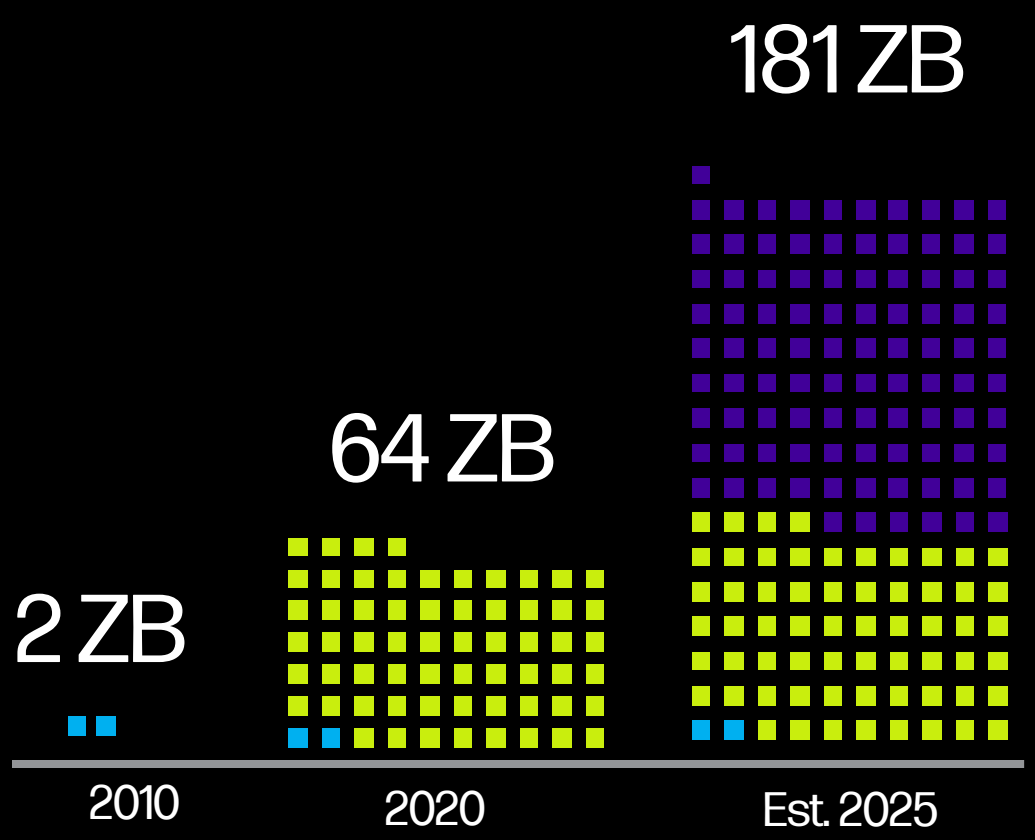
Structured

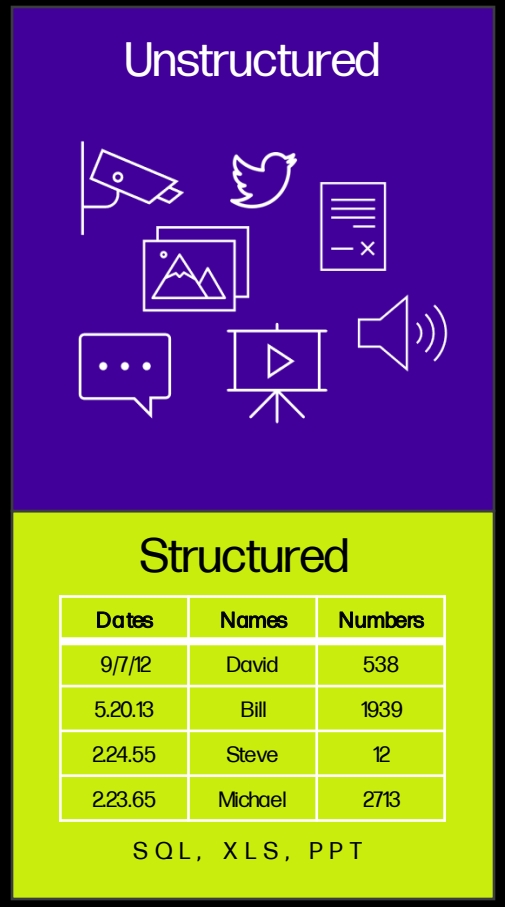
Dates	Names	Numbers
9/7/12	David	538
5.20.13	Bill	1939
2.24.55	Steve	12
2.23.65	Michael	2713

SQL, XLS, PPT

Explosion in Data Generation

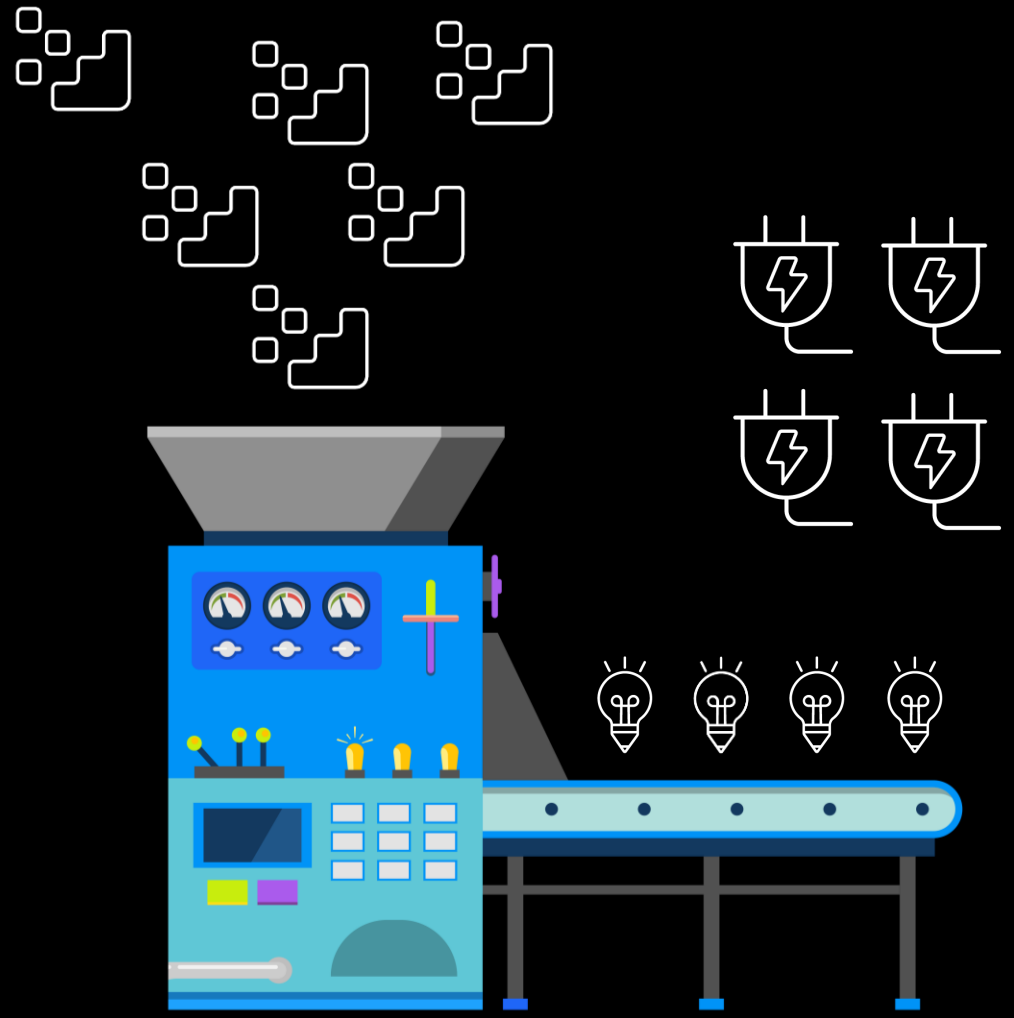
A trend that presents both challenge and opportunity



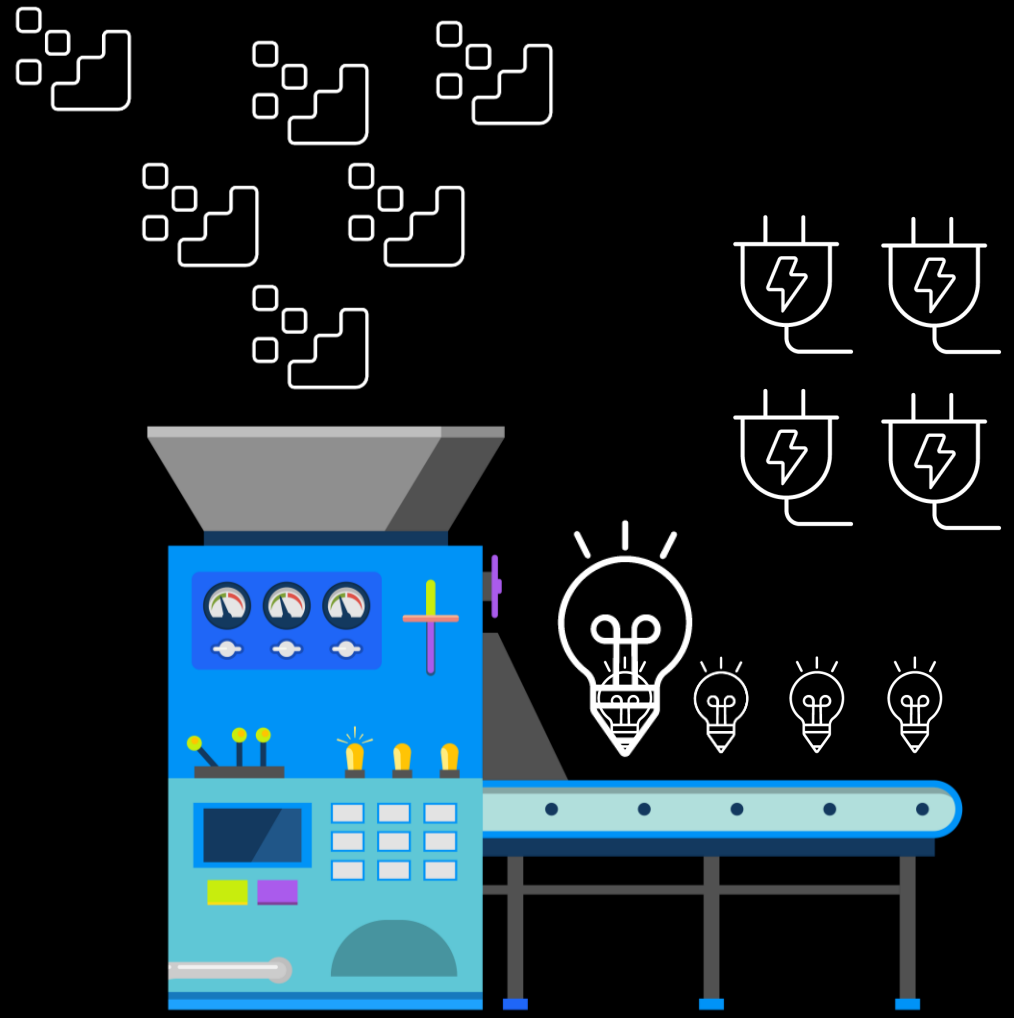


Growth by Data Type:

How does our factory process this volume?



How does our factory process this volume?



Priorities When Choosing Compute Resources

SPEED

How fast can I run the calculations?

SPEND

How much is this going to cost me?

SECURITY

Is my data safe?

Types of Compute Resources

CLIENT (PC)

PRO
- Data security
- Physical environment flexibility

CON
- Limited computing power
- Only available to individual users

EDGE

PRO
- Data security
- Powerful low latency compute

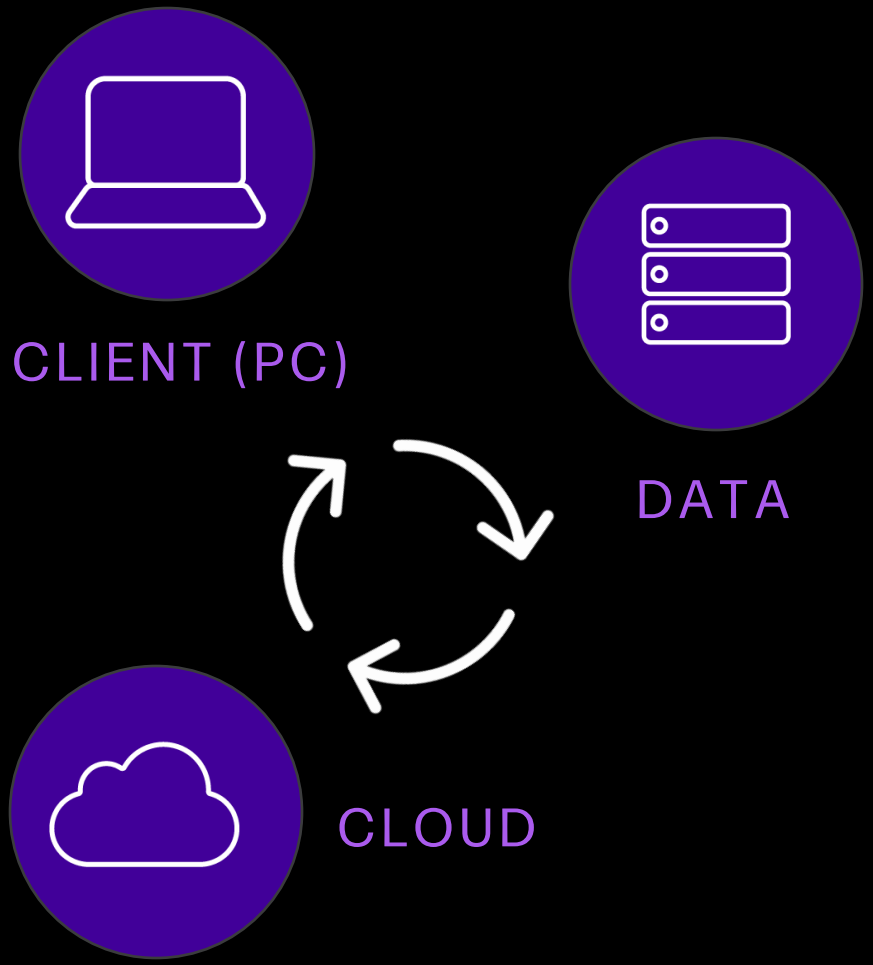
CON
- Physical environment limitations
- Inefficient when idle

CLOUD

PRO
- Powerful compute
- Minimal upfront cost

CON
- Data security risk
- Recurring cost

The Future of Data Processing



Hybrid Computing

enables data scientists to maximize the efficiencies of each type of computing resource, bringing together AI models quickly, cheaply, and safely.

Hybrid Computing Starts with Hardware

Z4 G5

Z6 G5

Z8 G5

Z8 Fury G5

Workflows

Z4 G5	Z6 G5	Z8 G5	Z8 Fury G5
• Data analysis • Structured data • Data preparation • Model prototyping	• Unstructured datasets (video, images, speech) • Machine learning • Streaming data analytics	• Streaming data analytics • Data visualization & simulation • Predictive modeling	• Deep learning, AI • Computer vision • Natural language processing • Data extract, transform, load (ETL)

Data

Smaller datasets & structured data → Larger datasets & unstructured data

Models

Less model complexity, fewer parameters → Increasing model complexity, more parameters

Hybrid Computing Starts with Hardware

ZBook Firefly G10 14”

ZBook Firefly G10 16”

ZBook Power G10 15.6”

ZBook Studio G10 16”

ZBook Fury G10 16”

Workflows

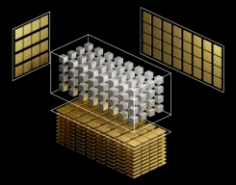
- Model building
- Remote data science connecting to Z Desktop via HP Anyware
- Exploratory data analysis
- Computer science and data science in education
- Data visualization
- Medium data analysis
- Light data science (ML inferencing)
- Heavy data analysis
- Medium data science (ML inferencing and training; DL inferencing)

Data — Smaller datasets/structured data → Larger datasets/structured data

Models — Model building → ML inferencing → ML inferencing/training & DL inferencing

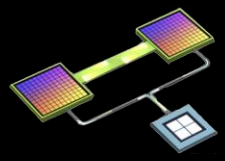
NVIDIA A800 40GB Active

Based on the NVIDIA Ampere architecture

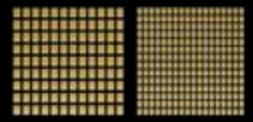


3rd Gen Tensor Cores
Accelerated AI

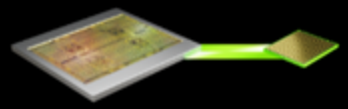
NVIDIA AI Enterprise
3-year subscription Included



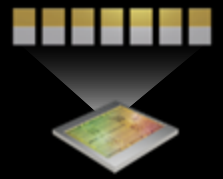
3rd Gen NVLink
Scalable performance



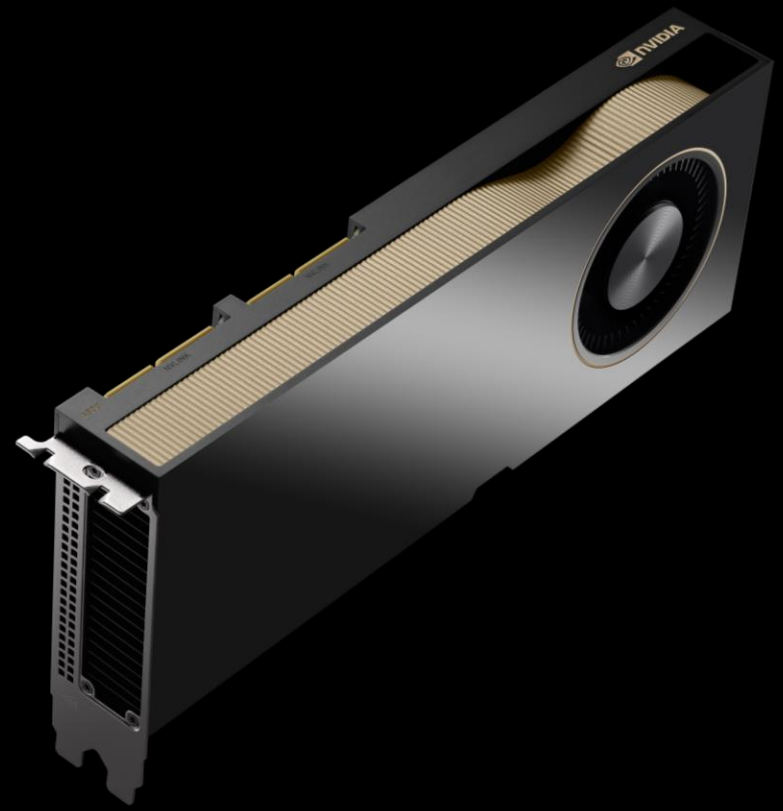
40GB HBM2 Memory
For large models & datasets



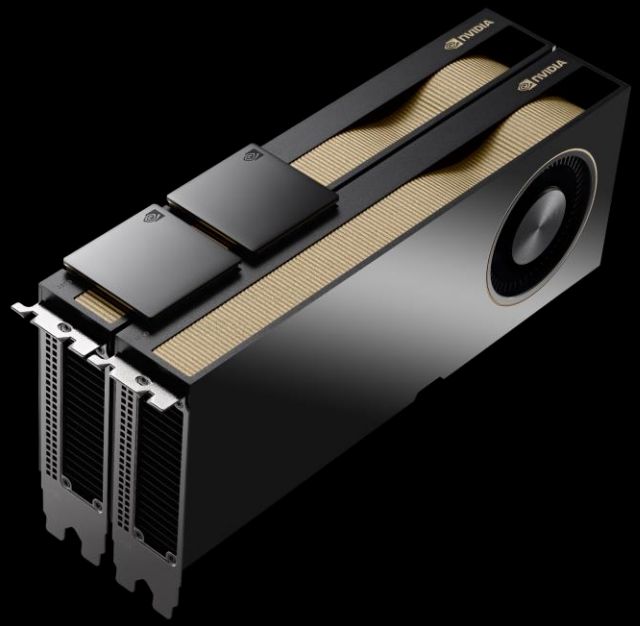
1.5 TB/s Memory Bandwidth
Ultra-fast data transfers to and from GPU memory



Multi-Instance GPU (MIG)
Optimize multi-instance workloads



A800 Specs



Dual NVIDIA A800 40GB Active w/NVLink

GPU Memory	40GB HBM2
Memory Interface	5120-bit
Memory Bandwidth	1555.2 GB/s
CUDA Cores	6912
Tensor Cores	432
Double-Precision Performance	9.7 TFLOPS
Single-Precision Performance	19.5 TFLOPS
Peak Tensor Performance[1]	623.8 TFLOPS
Multi-Instance GPU	Up to 7 MIG
NVIDIA NVLink	Yes
NVLink Bandwidth	400GB/s
System Interface	PCIe 4.0 x 16
Power Consumption	240W
Thermal	Active
Form Factor	4.4" H x 10.5" L - Dual Slot
Display Outputs	-

1. FP16 matrix multiply with FP16 or FP32 accumulate
2. Requires a companion GPU to support display-out

When to Use A800 / RTX 6000 Ada Generation

	DL / ML				Generative AI				Compute/HPC		Visual Computing[1]			
	Training < 40GB Single GPU	Training > 40GB Single GPU	Inference	FP8[2] data format support (training /inference)	Fine Tuning	Fine Tuning (FP8)	LLM Inference	Image Generation	FP64	FP32	Graphics	Rendering	Omniverse	NVLink Support
RTX 6000 Ada Generation	●	●	●	●	●	●	●	●		●	●	●	●	
A800 40GB Active	●		●		●		●	●	●	●	●	●		

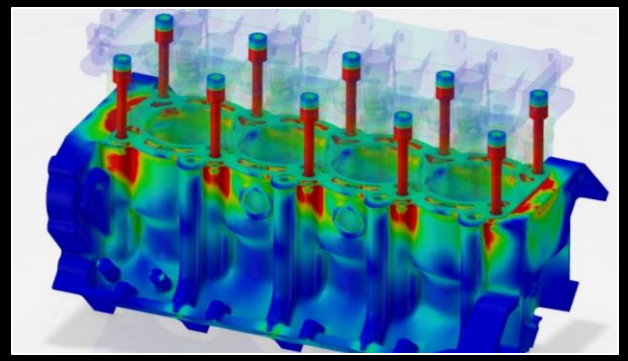
● Best Performance

● Good Performance

1. A800 40GB Active requires a companion GPU for graphics display capability. Some applications may not run without direct display from the GPU
2. FP8 Data format support only on RTX 6000 Ada Generation
3. 2x A800 with NVLink can provide 80GB of combined GPU memory for applications, frameworks, and toolkits that support NVLink

A800 Use Cases

Engineering Simulation/CAE



MFG, Auto, Aerospace

Energy & Geosciences

Exploration, Seismic Analysis, Nuclear Research

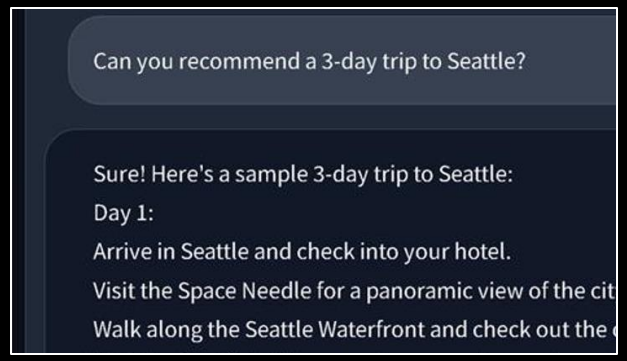
AI Training & Development

Model training, fine tuning, experimentation

Data Science

Data Prep, Model Training, Evaluation

Generative AI



LLMs, RAG

High Performance Compute (HPC)

Simulations, Molecular Dynamics, Physics

What can AI workstations do for my business?

"With the data science workstations, we're able to get high accuracy with the models-at least 90 percent. This helps us plan our cargo freights better than we ever could before."

Tassio Carvalho
Head of the Center for Machine Learning and Artificial Intelligence at American Airlines

"My typical benefit (using an HP Z8 workstation) is a speed up of about 1 order of magnitude with respect to my previous workflow using remote computer clusters. I'm able to train neural nets in seven minutes when it was taking an hour on a (remote) computer cluster, and inconceivable on a laptop."

Raphael Attié
Solar astonomer at NASA Goddard

"We can train color models and run experiments that used to take us five days in just 36 hours. More importantly, with the giant GPU memory on the Z8, we can tune hyperparameters to run bigger batch sizes, which translates to improved results."

Artem Legotin
Machine Learning Engineer, neural.love

Software Brings AI Projects Together

Z by HP machines come with all the software you need to complete data science projects, right out of the box

Software Tools

Get the most popular tools for analyzing data and creating powerful models. Enjoy seamless management of package updates and dependencies so that your environment is always ready-to-use.

Developer Utilities

Easily create and deploy your data science applications while managing your data science tools and models.

Enhanced Cloud Experience

Interact seamlessly with your main cloud environment with cloud command line access.

Ubuntu. Certified and Preinstalled.

We work closely with Canonical to extensively test and certify the latest version of Ubuntu OS so it performs at its best.

Microsoft WSL 2 Preinstalled

Experience the best of both Windows and Ubuntu with Microsoft Windows Subsystem for Linux 2. WSL 2 is a simple and fast way to run Linux directly in Windows.

AI Studio Puts the Pieces Together

One platform to connect people, data, tools, and compute to accelerate model development

CENTRALIZE

All your data, tools, team members, and projects on one platform.

COLLABORATE

Easily access and share data, templates, and experiments with your team

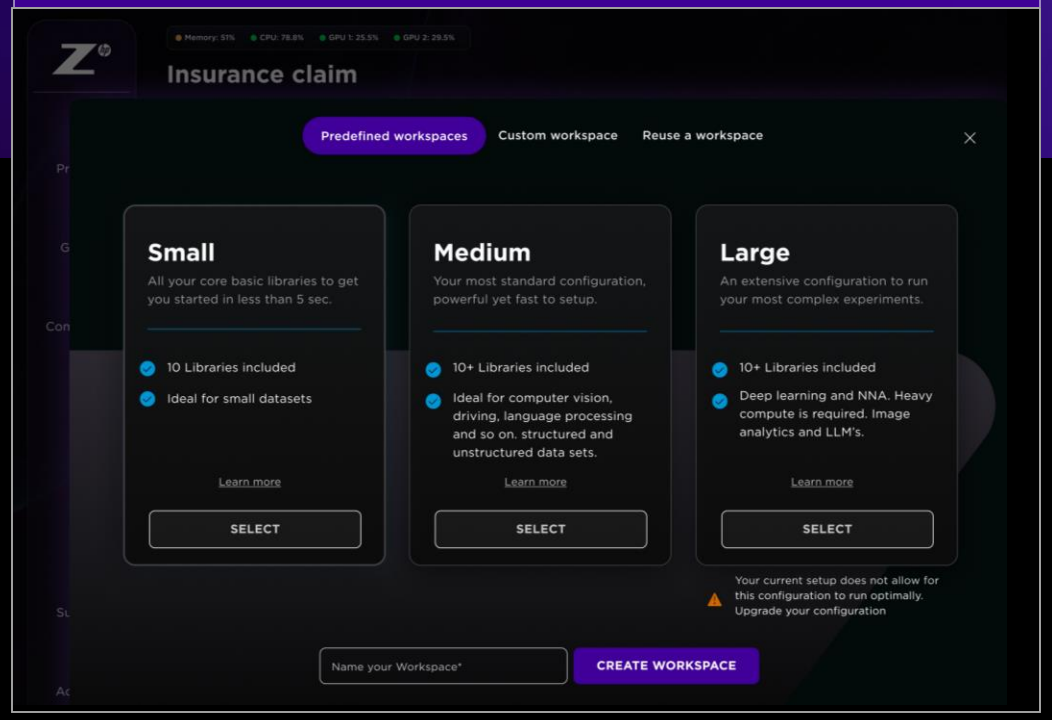
ACCELERATE

Speed up AI model delivery through operational efficiency

Instant Environment Creation

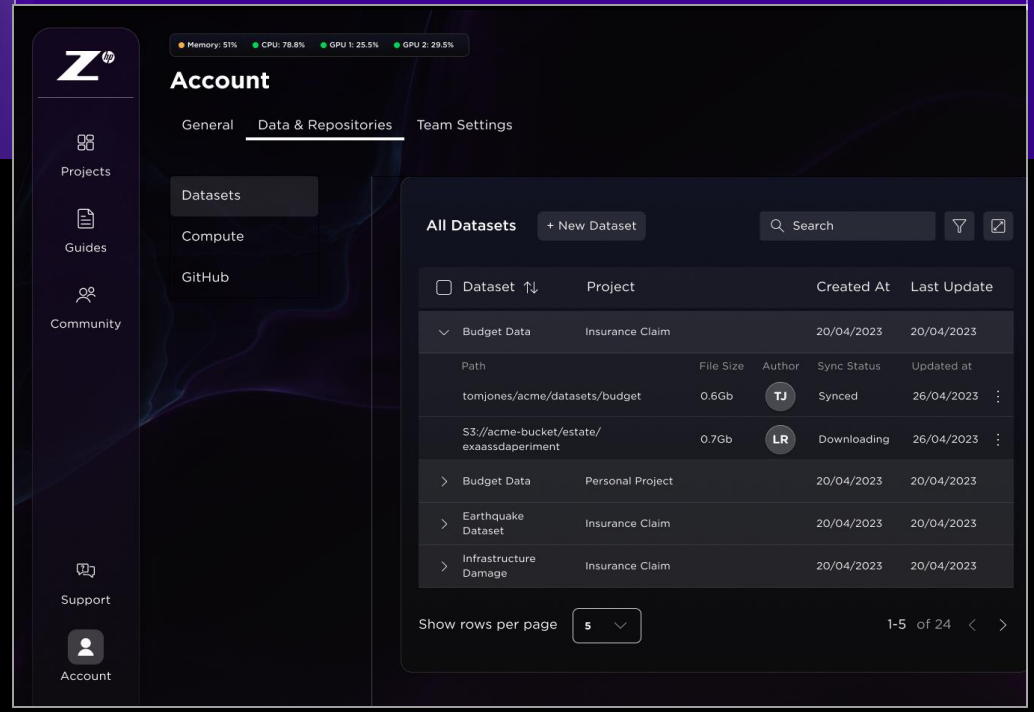
Flexible Project Setup

Customize, reuse, or use a predefined workspace to accelerate project setup.



Account Management

Manage all access to data, datasets, projects, and GitHub repositories from the administrator account.

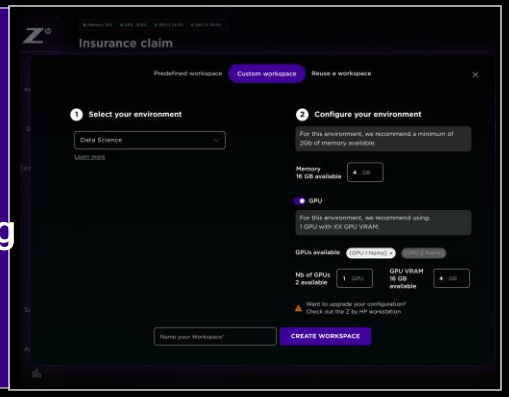


*Cloud computing requires an AWS cloud account.

Easy Team Management & Collaboration

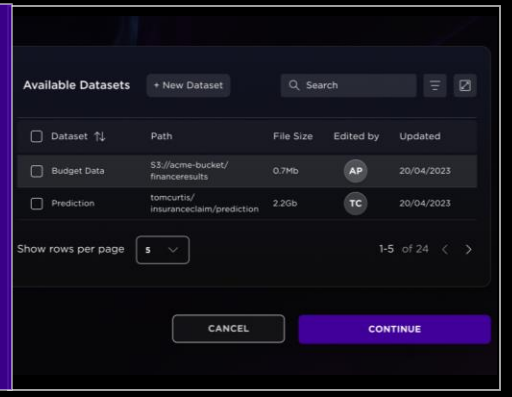
Shared Workspaces

Leverage shared workspaces for consistency of libraries and packages among team members.



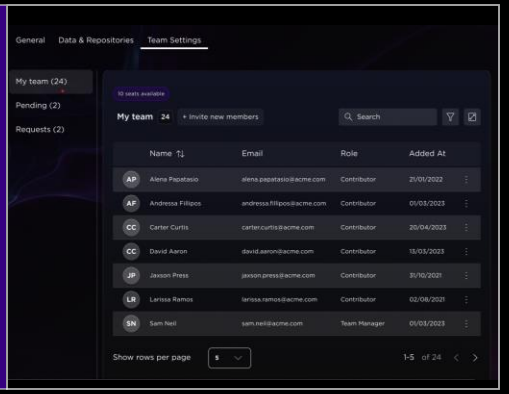
Manage Data Access

Use the platform to access your data on-prem or from the cloud and manage access to those datasets.*



Add Team Members

Invite team members to the platform to collaborate on projects in real-time.

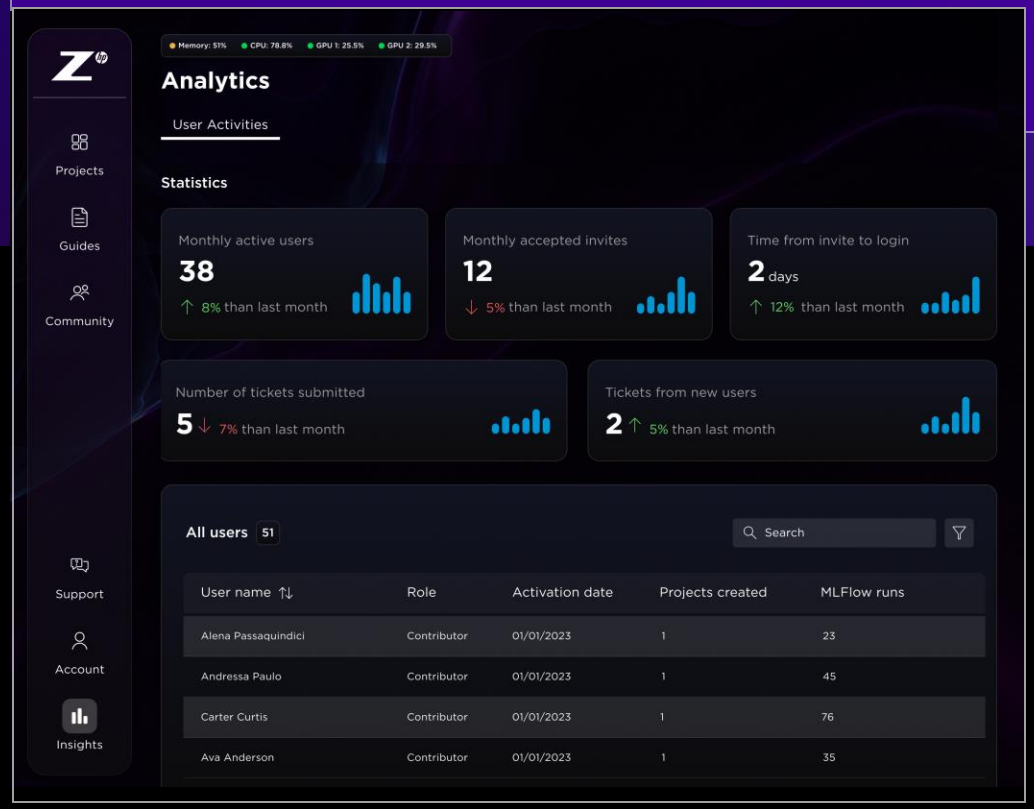


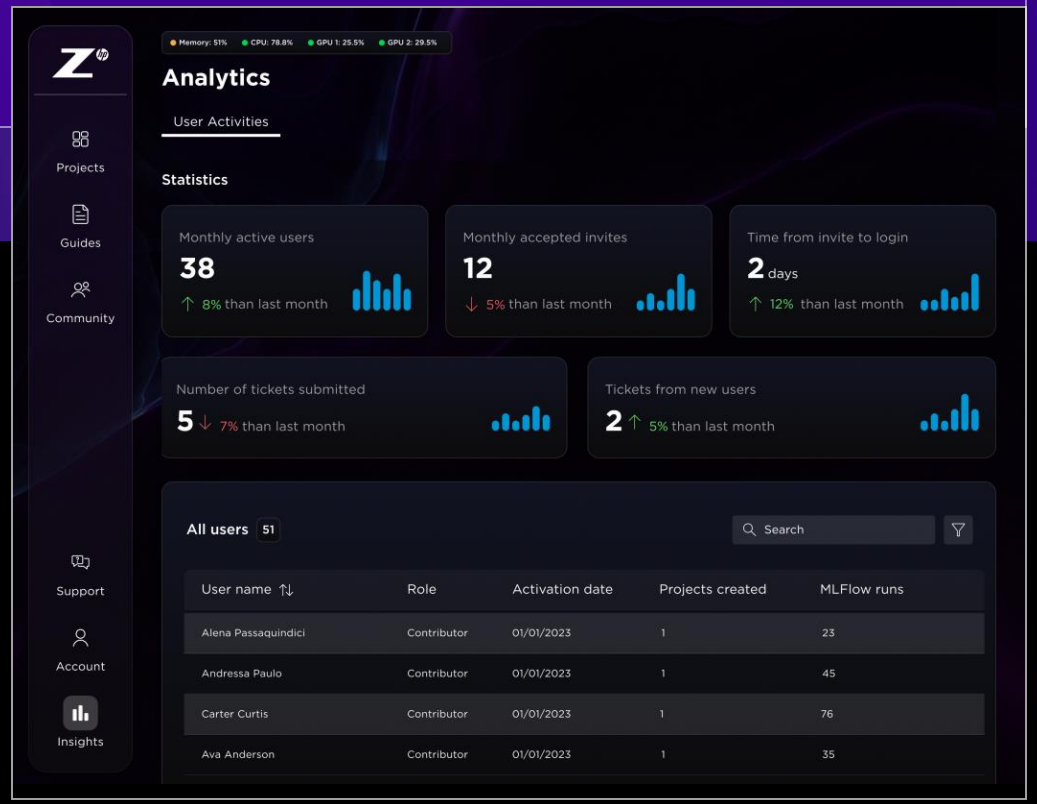
*Cloud computing requires an AWS cloud account.

Powerful Analysis Tools

Analytics Dashboard

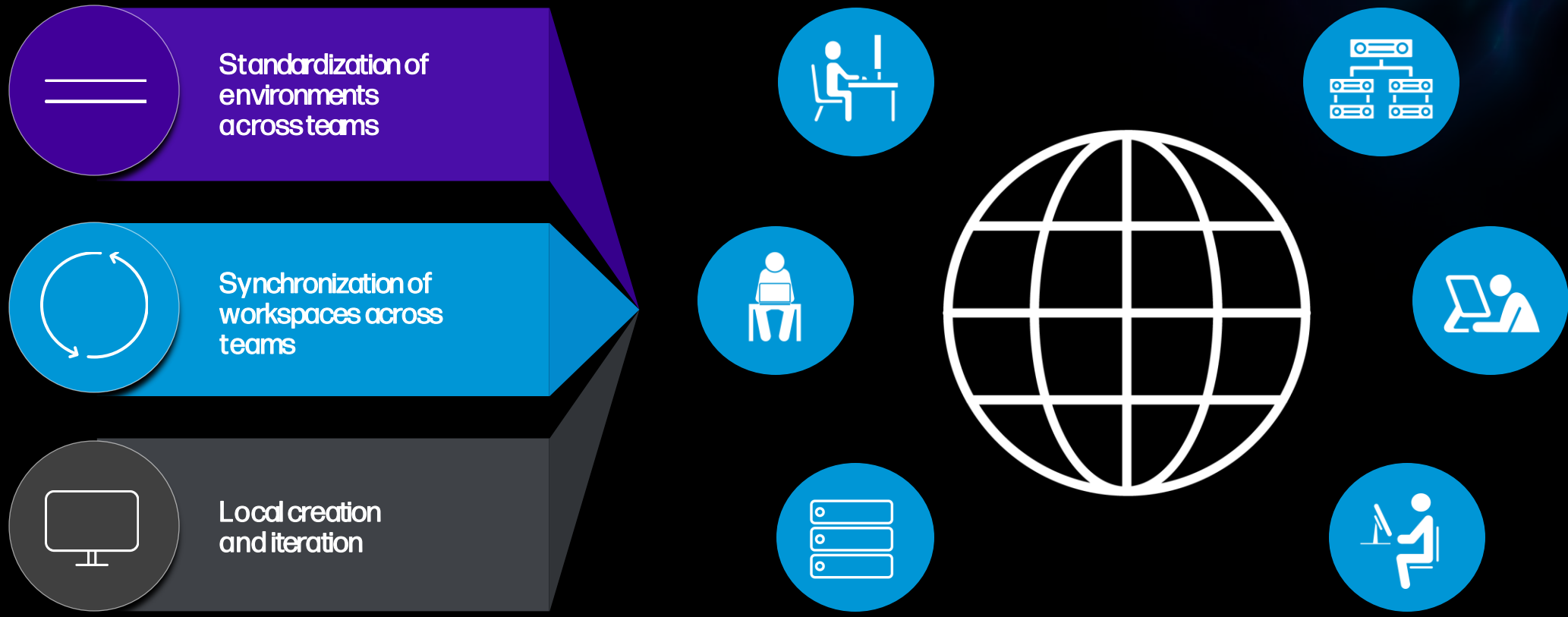
Ensure optimization of your resources with a comprehensive overview of all your compute in use.





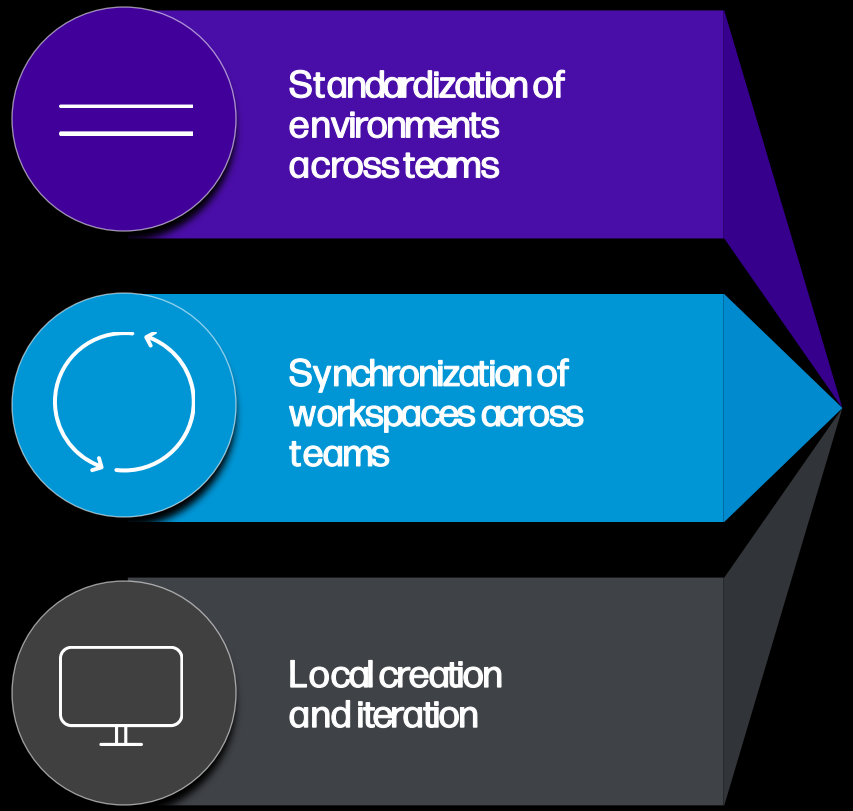
Transforming your workflow

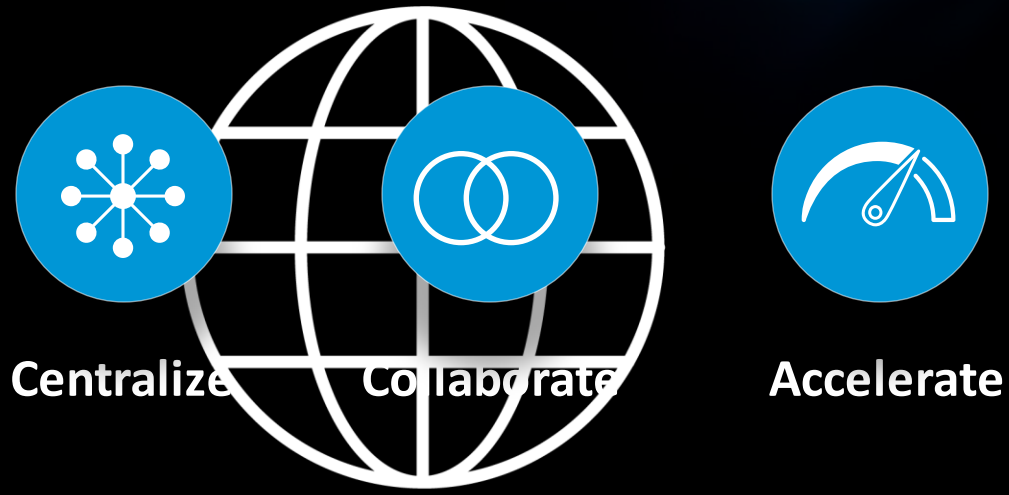
No matter where the works starts, AI Studio brings it all together



Transforming your workflow

No matter where the works starts, AI Studio brings it all together





Public. For use with general public.