WEKA for GenAI

Shimon Ben David – CTO

A Majority of Enterprises Are Now Investing In Generative AI

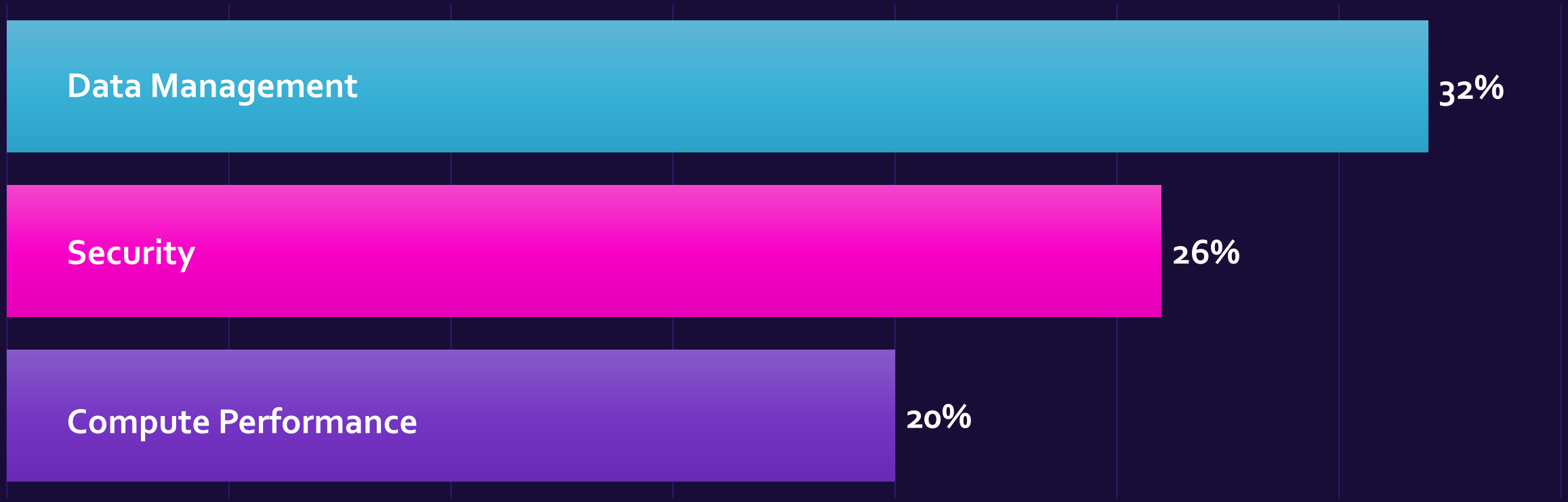
Source: S&P Global Market Intelligence 2024 Trends in Data, AI & Analytics, December 2023

By 2027, 90% of enterprises will deploy generative AI models and applications —up from less than 5% in 2023.

Source: Gartner Inc, Dec 2023 https://www.gartner.com/document/code/791364

K E Y F I N D I N G :

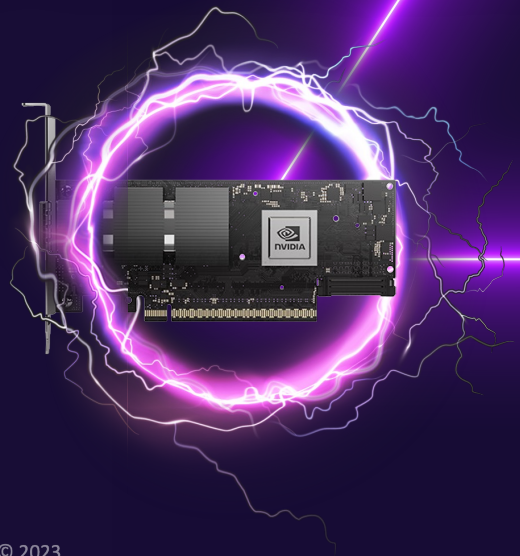
Top technical inhibitors to AI/ML success



WEKA

The Infrastructure Triangle

Compute
Next generation GPU Accelerators



Networking
Modern 400GbE/IB fabrics
(800Gb/s Switches)

"Vintage" Storage

Antiquated Storage
- 30-year-old protocol (NFS)
- Unable to saturate modern networks
- Lots of small files overwhelming MDS
- etc.

WEKA

GPU Acceleration
1000x vs Traditional CPU

2008

First Top500 supercomputer

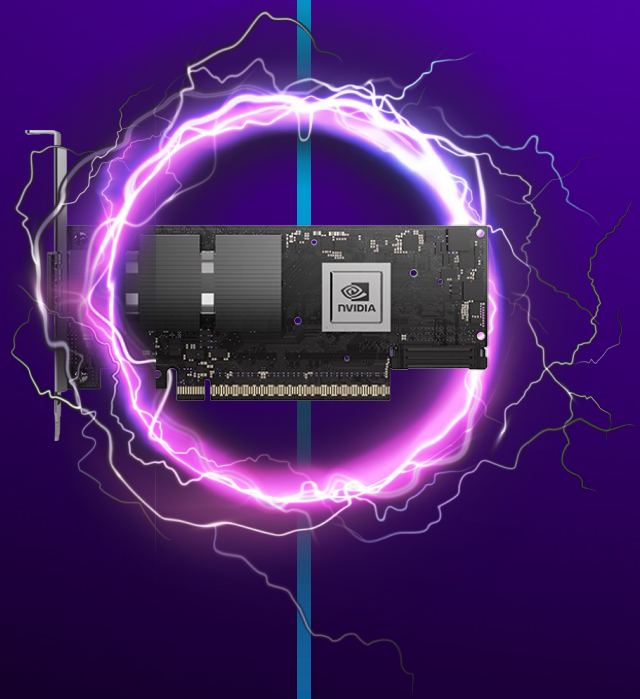
- 10,400 Rack Units
- 1 PetaFLOP

2022

NVIDIA DGX-H100

- 8 Rack Units
- 32 PetaFLOP

WEKA



Modern Networking
80x Improvement vs legacy

2002

10GbE IEEE Standard

- Full duplex point-to-point links
- 10 gigabits per second

2022

NVIDIA ConnectX-7 400Gb

- NDR InfiniBand and 400Gb Ethernet
- 400 gigabits per second, with switch back hauls of 800Gb/s

WEKA

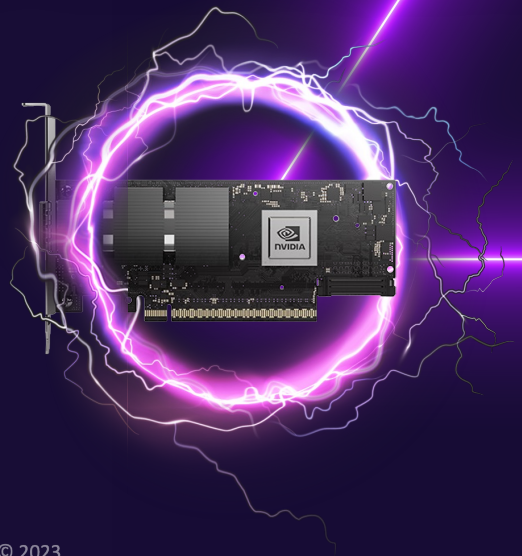
“Vintage” Storage

Traditional PFS
All Flash NAS
Software Defined ‘NFS’

WEKA

The Infrastructure Triangle

Compute
Next-generation GPU Accelerators



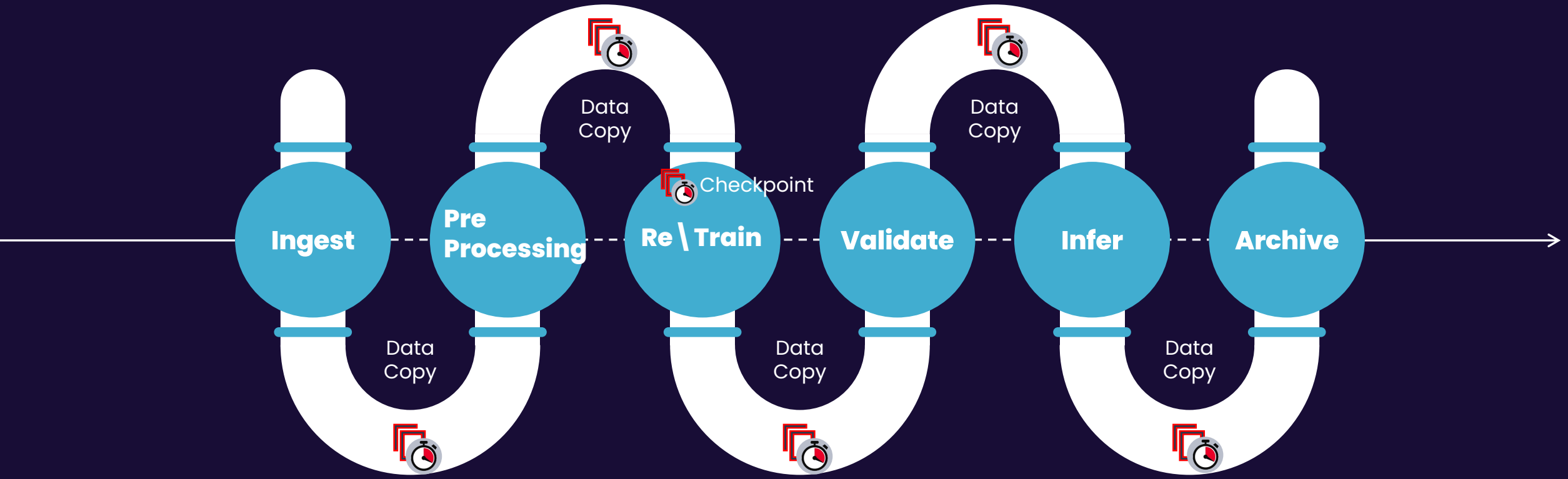
Networking
Modern 400GbE/IB fabrics
(800Gb/s Switches)

WEKA Data Platform
Modern Data Infrastructure

WEKA

AI Pipelines Break Traditional Data Silos

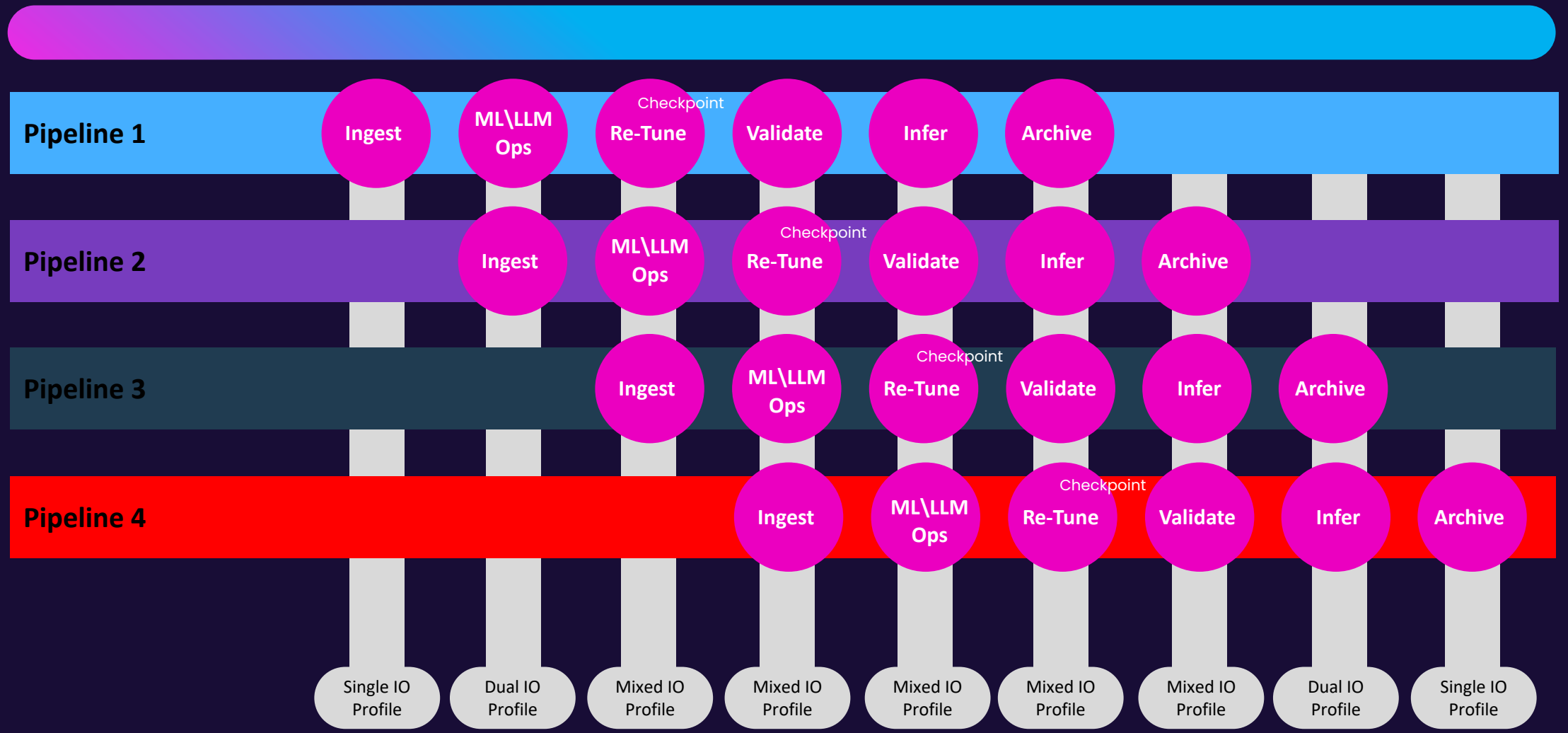
Energy and time wasted on keeping copies



Massive concurrency, Write throughput	Annotation, index, search, cloud bursting	Massive read throughput, massive writes for checkpointing	Massive bandwidth for large number of streams replay	Low latency access	Lifecycle mgmt., versioning, reproducibility

WEKA

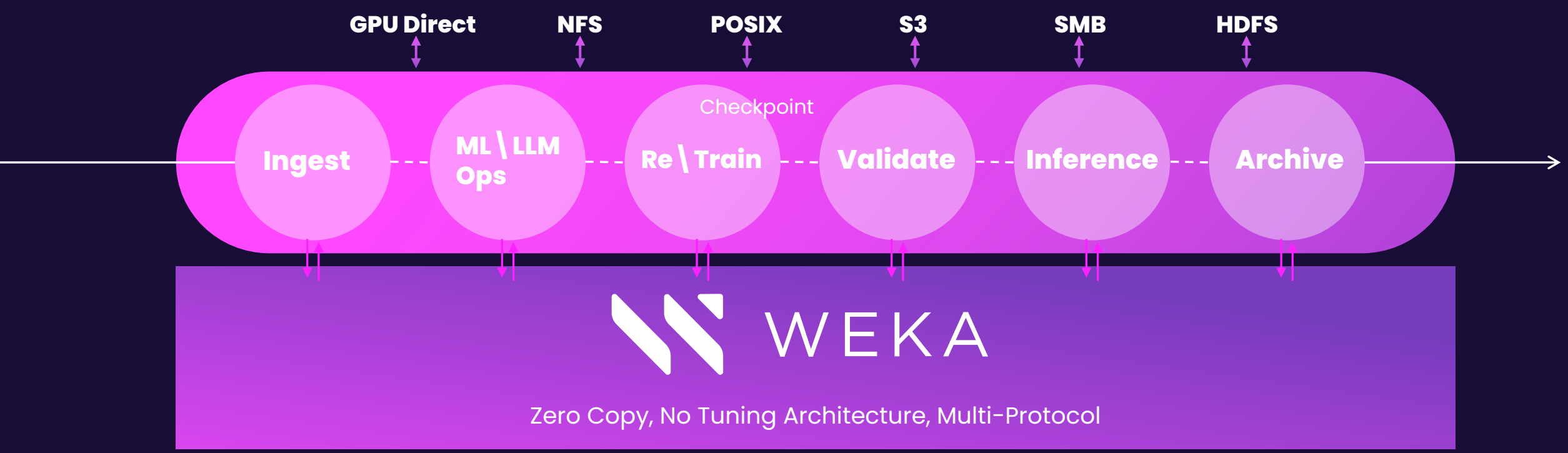
AI Data Pipeline: multiple pipelines heating storage No tuning is possible



WEKA

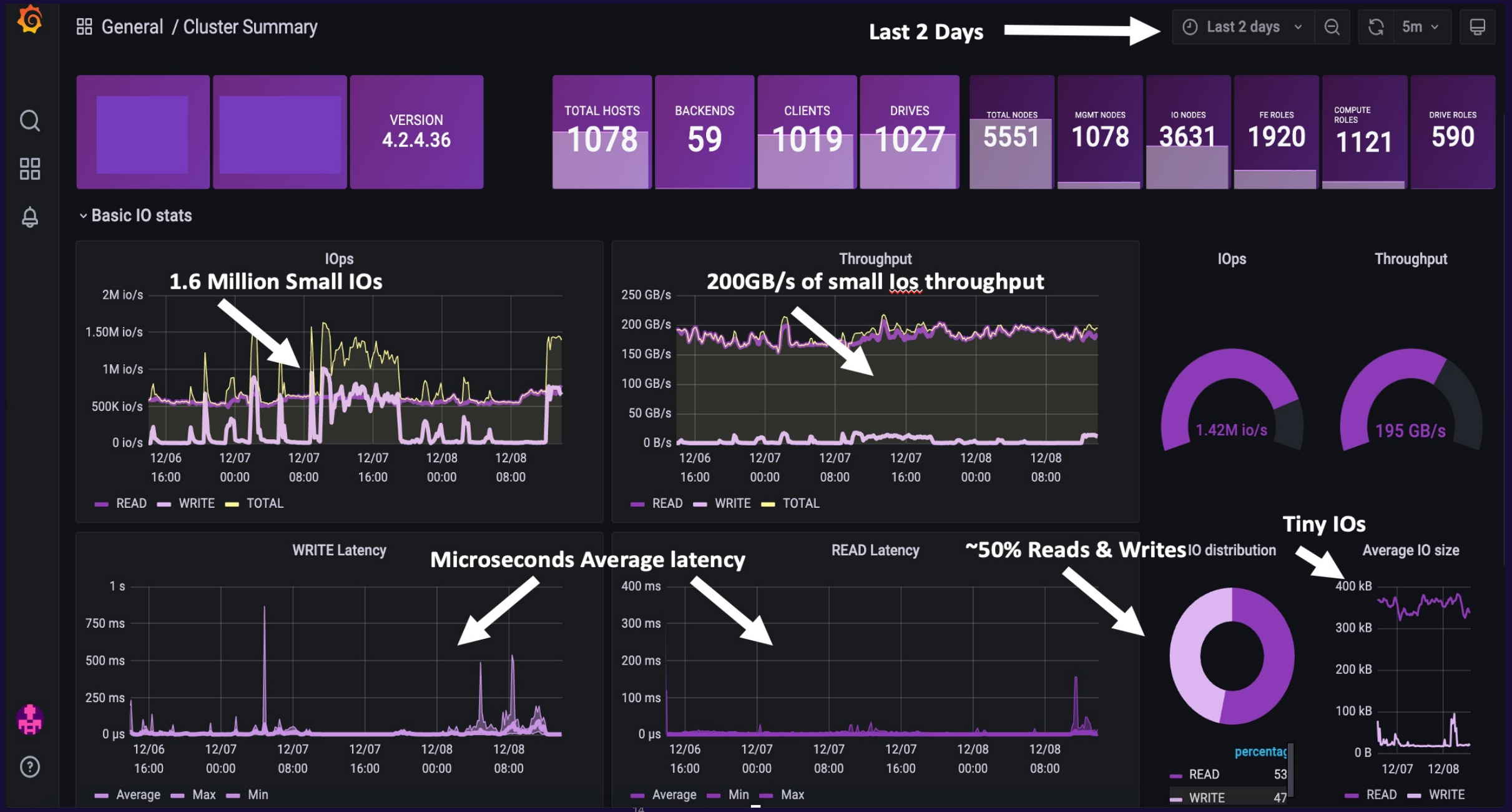
The WEKA Data Platform

No copying, faster performance leads to x10–100 improvement in time to epoch



Production Weka AI Systems Measured IO Patterns

AI customer #1 IO Pattern – Millions of Tiny IOs Reads / Writes



Checkpointing Example

Every 30-60 minutes
1000 GPUs X 80GB < 60 seconds
1.3TB/s Writes to shared storage

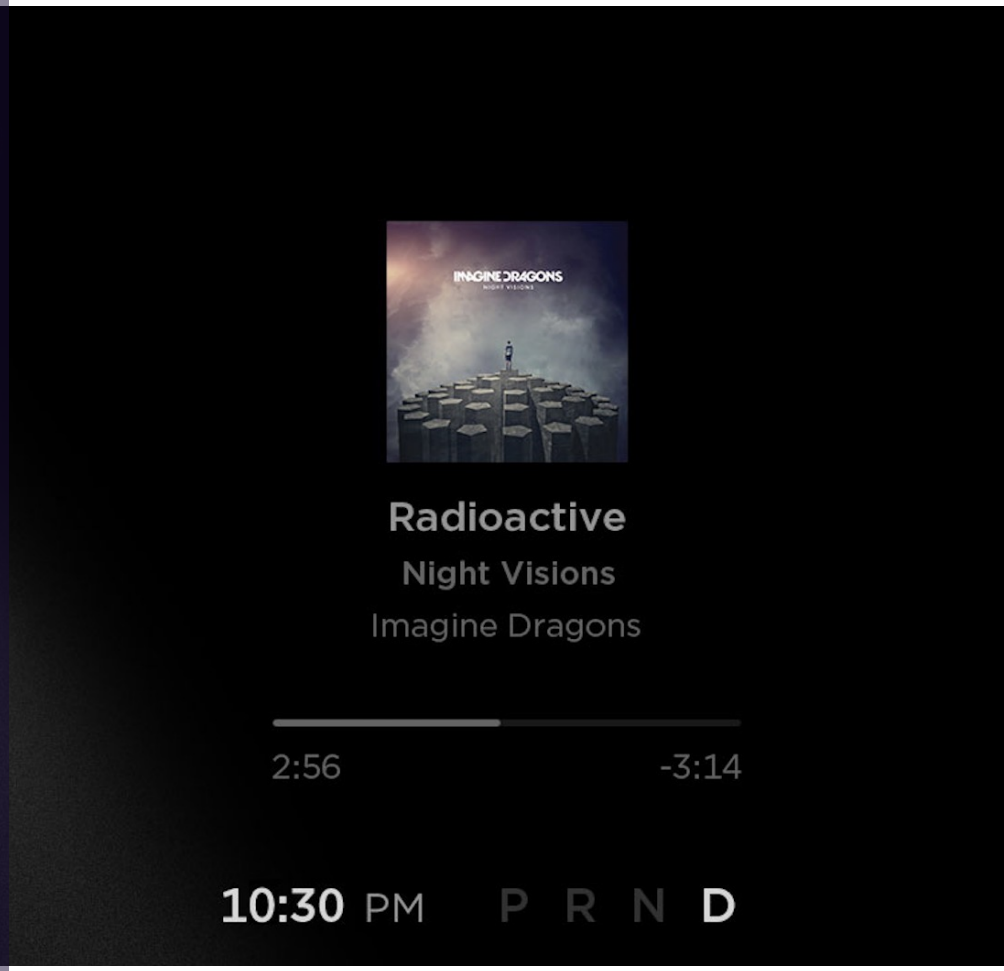
Inferencing Example

Model repositories of 100s TBs
Time to Load models to inference server
Time to output GenAI Artifacts
Accelerate VectorDB & Embedding

‘Impossible’ Workflows powered by **WEKA**

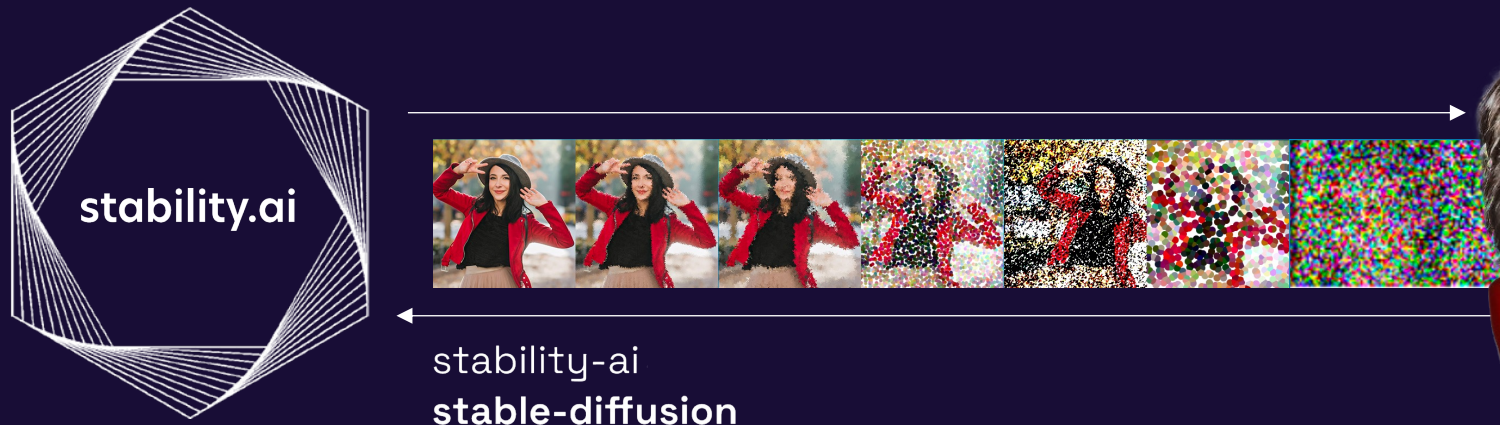
WEKA

Advanced Autopilot. 70x Faster.



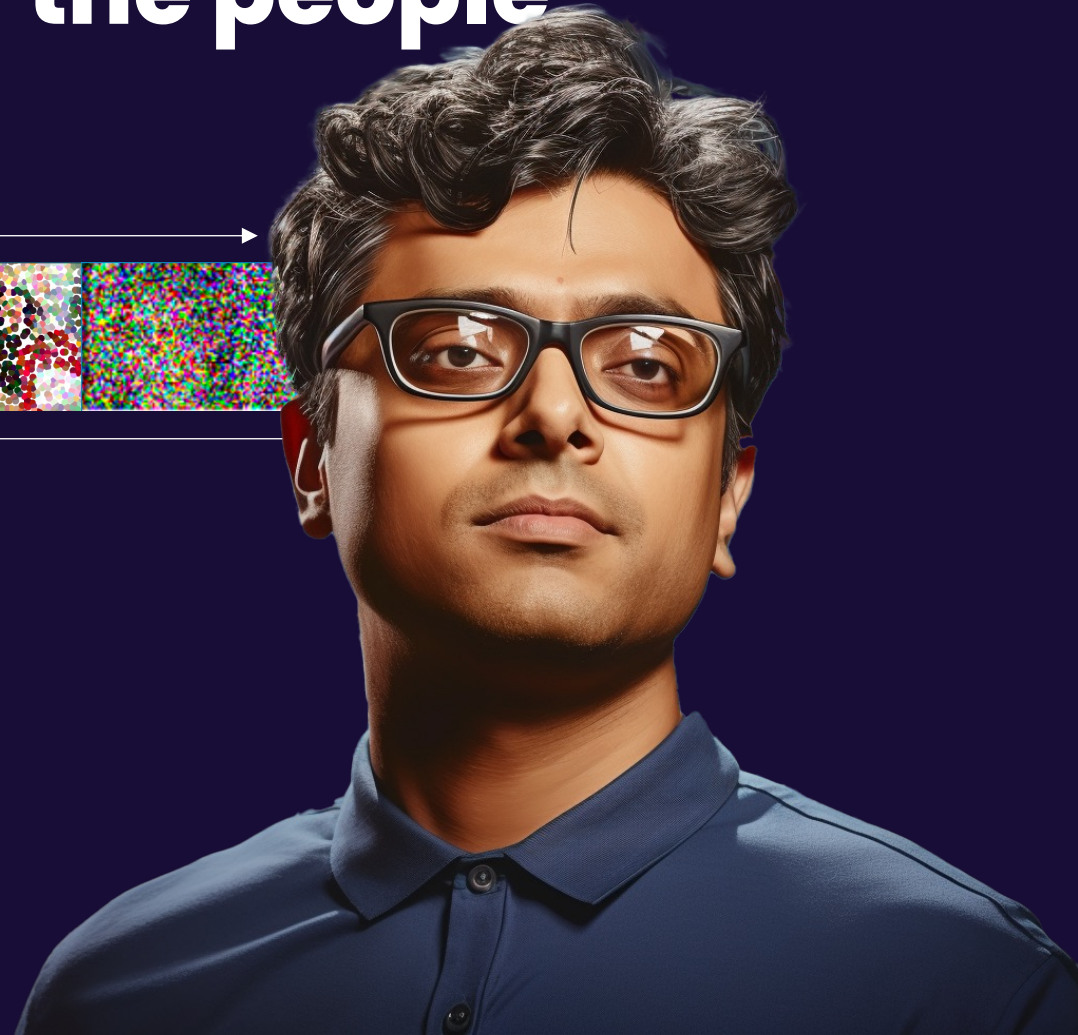
WEKA

AI by the people, for the people



stability-ai
stable-diffusion

A latent text-to-image diffusion model capable of generating photo-realistic images given any text input



- **Delivered 80%+ cost savings**
- **Increased GPU Utilization by 2.5x vs Lustre**
- **Increased data transfer 40x vs Lustre**
- **Running in Converged on GPU servers in AWS**

One Year.
Two Vaccines.

WEKA

One Planet.
Protected.

SLAC NATIONAL ACCELERATOR LABORATORY

U.S. DEPARTMENT OF ENERGY

WEKA

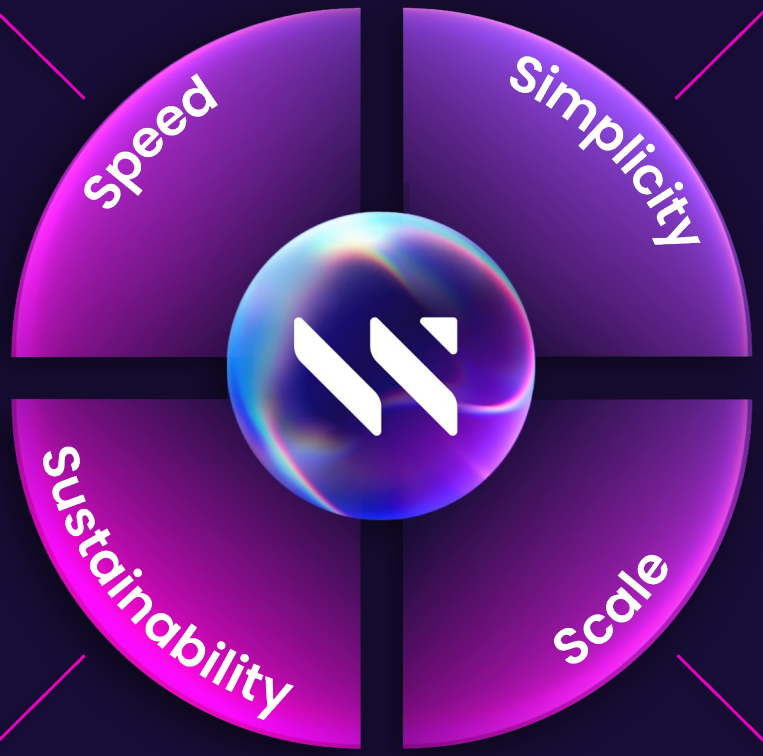
One Architecture Delivers on FOUR Promises

Mindbending Speed

Deliver unbeatable file and object performance for your most demanding applications supporting high I/O, low latency, small files, and mixed workloads with no tuning.

Seductive Simplicity

Eliminate the complexity and compromises of traditional data infrastructure with a single, easy-to-use data platform that eliminates storage silos across on-premises and the cloud.



Effortless Sustainability

Lower energy consumption and reduce the resulting carbon emissions by cutting data pipeline idle time, extending the usable life of your hardware, and moving workloads to the cloud.

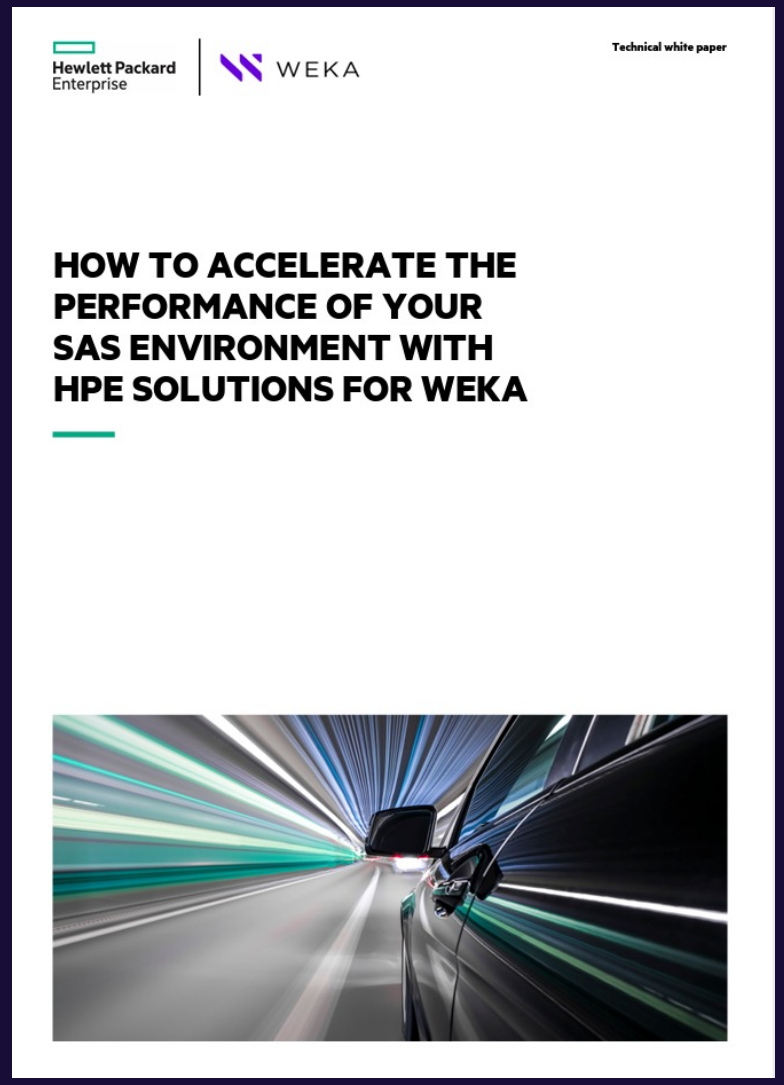
Infinite Scale

Scale your compute and storage independently and linearly on-premises or in the cloud with WekaFS to handle 10s of millions or even billions of files of all data types and data sizes.

WEKA

Why HPE with weka?

- HPE is an early investor in WEKA (relationship since 2017) through the HPE Pathfinder Program with executive level relationships in place (Board level representation). HPE has participated in multiple funding rounds.
- WEKA is a Strategic Partner in the HPE Complete, the one-stop shop for validated HPE and third-party partner end-to-end infrastructure solutions.
- WEKA has won the HPE Technical Partner Program Momentum Partner of the year.
- Long-standing engineering and support relationships between HPE and WEKA have led to insights and domain-specific work in areas from AI for medical imaging to analytics.



Thank You