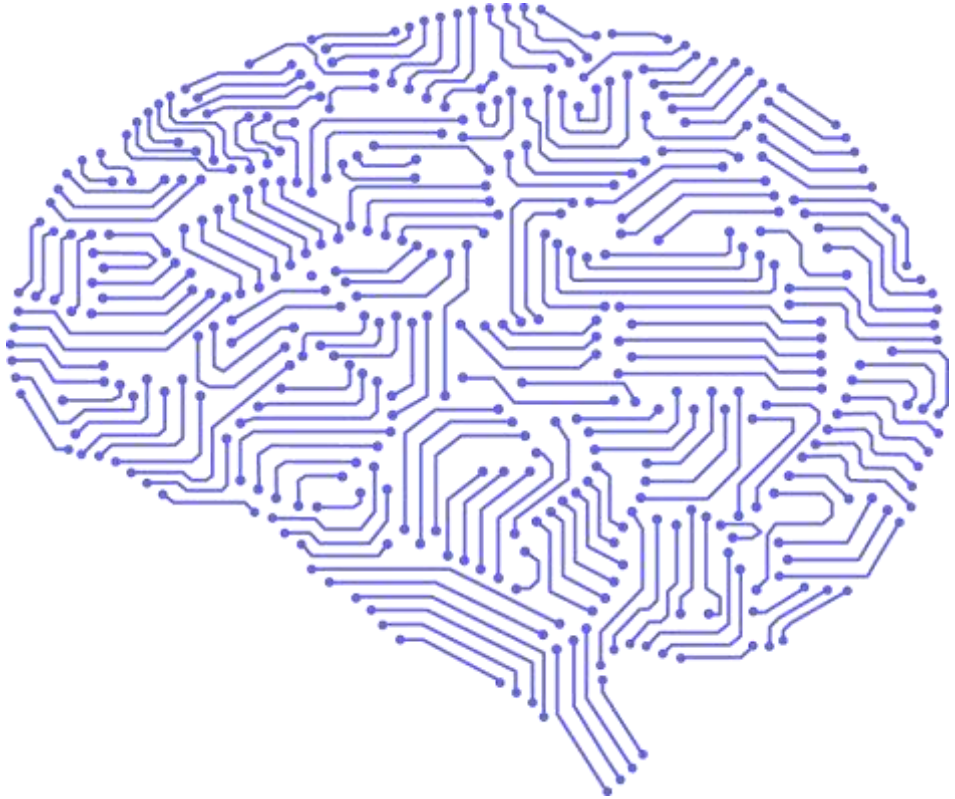
cnvrg.io

Continuous training and deployment of AI



John Palazza
Vice President of Worldwide Sales

Getting value from AI is hard

AI projects take too long to deliver value, if they're delivered at all

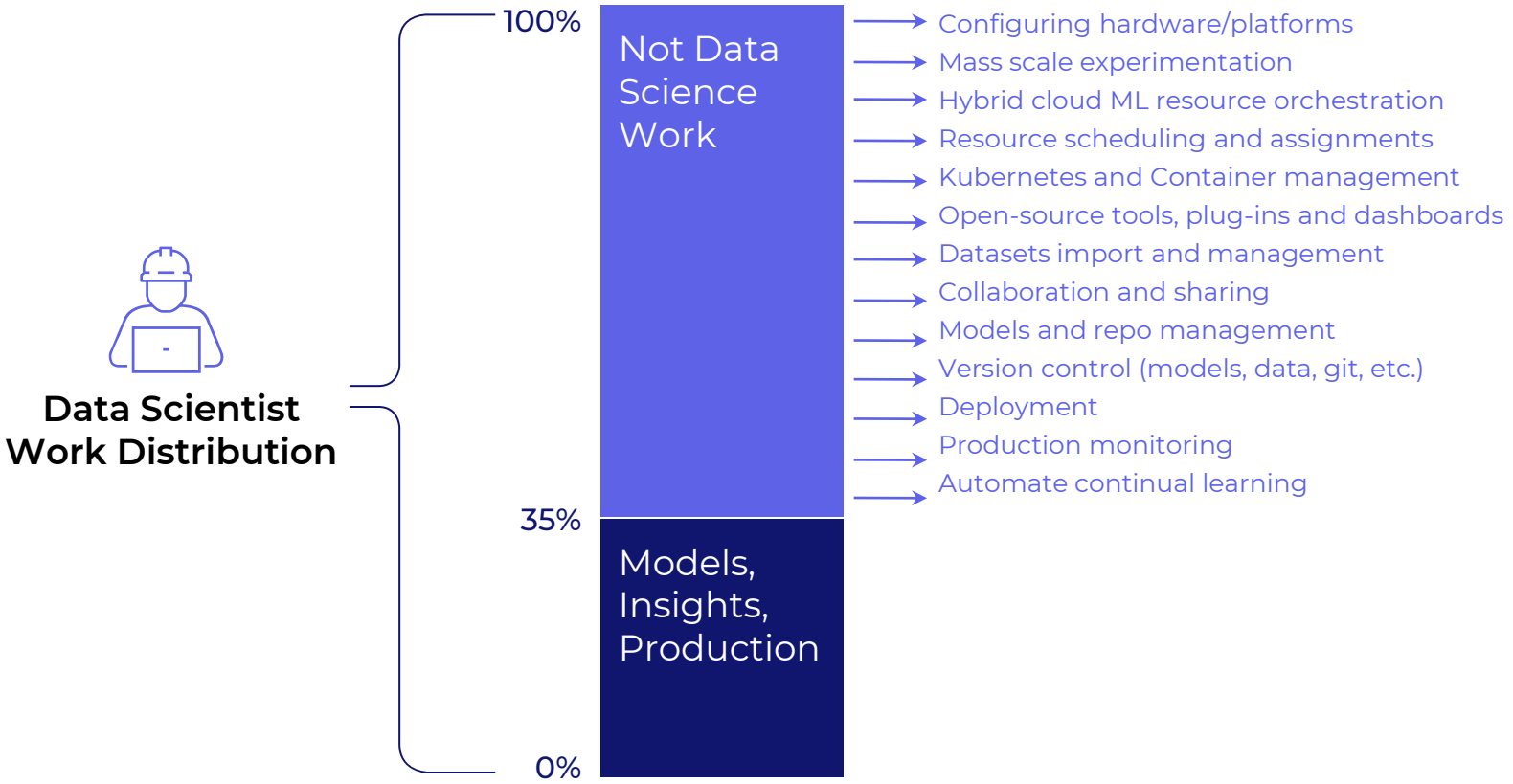
According to Gartner research, a bare majority of AI projects eventually move beyond the lab into production

It takes an average of

9 months

to get out of the lab

Data scientists spend too much time on incidental tasks



50%
OPEX
WASTE

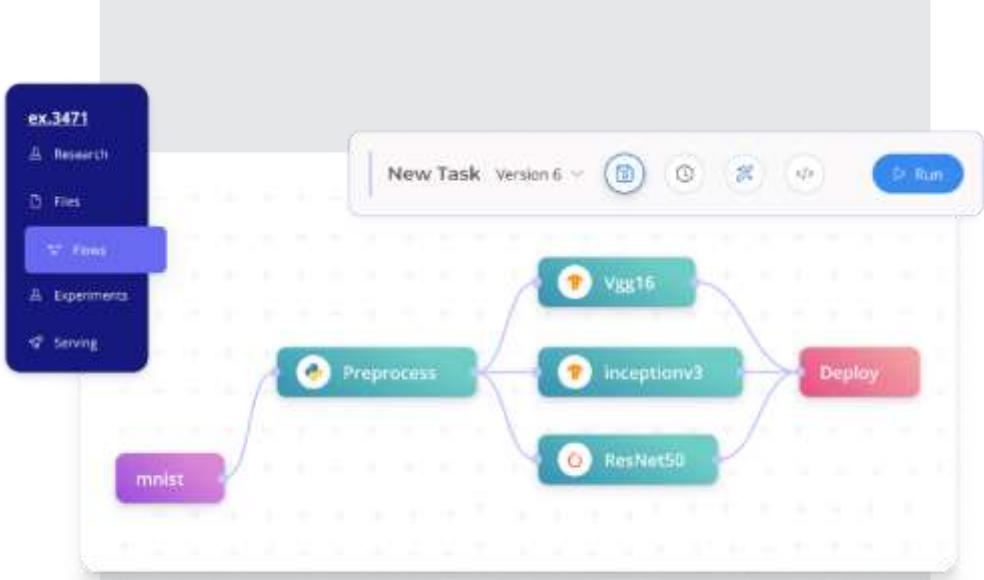
2X
SLOWER TIME
TO MARKET

30%
REVENUES
LOST

cnvrg.io Overview

Built by data scientists for developers of AI applications

- A platform to automate the **continuous training and deployment** of AI and ML models.
- Manages the **entire lifecycle**: data preprocessing, experimentation, training, testing, versioning, deployment, monitoring, and automatic retraining.
- Enables developers to train and deploy on **any infrastructure at scale**
- cnvrg.io **Metacloud** is the cnvrg.io platform offered as a **managed service**

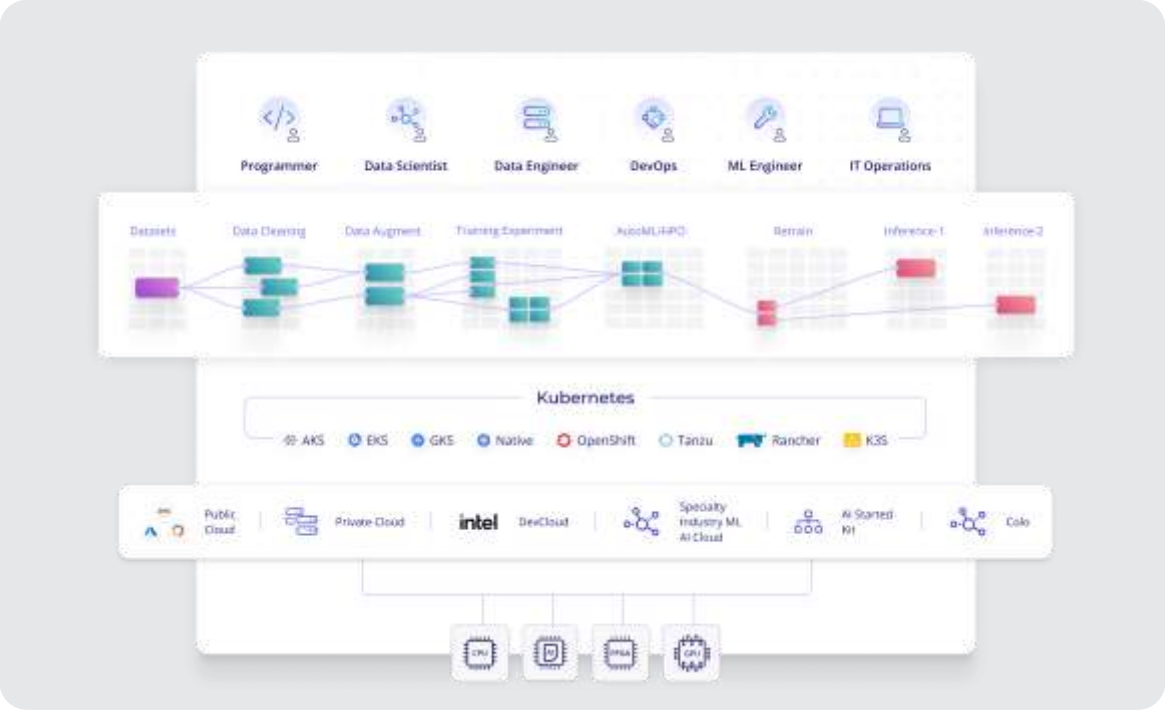


Benefits

- Up to **10x increase in productivity**
- Up to **5x faster model training**
- Up to **50% increase in compute utilization**

cnvrg.io: Operating System for AI

Everything needed to build and deploy AI on any infrastructure



Control Plane
Management layer for datasets, model code, jobs, model performance, cluster and resource statistics

AI Library
Package manager for algorithms and data components, with Git integration for adding your own repositories

Pipelines
Drag-and-drop interface for building end-to-end ML pipelines

Orchestration and Scheduling
Kubernetes-based meta-scheduler for orchestration, scheduling, and scaling across clusters

Compute and Storage
Connect your own compute and storage, or choose partner-provided resources from our marketplace

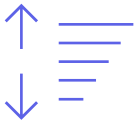
NEW

cnvrg.io Metacloud

Our AI infrastructure stack as a managed service

Build frictionless machine learning pipelines in just a few clicks

Connect any data source and control your datasets in one place



Use and customize our library of models, or create your own

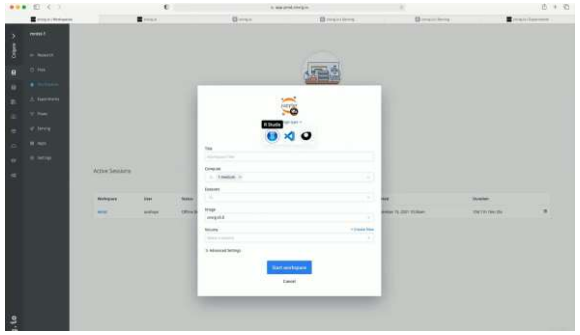
Manage multiple experiments in parallel and track performance

Instantly deploy AI models to production

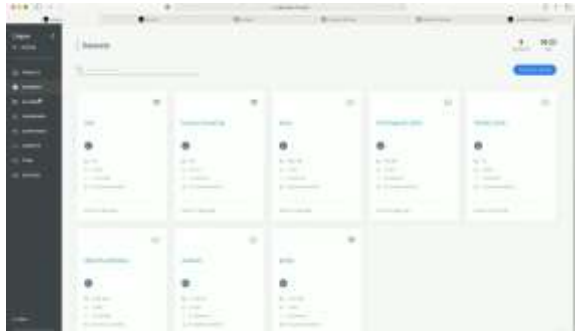
- One-click setup
- Centralized artifact repository
- Version control of data, code, and artifacts
- Heterogeneous pipeline automation
- Continuous training and deployment
- Automatic retraining based on model/data behavior
- Performance and health metrics for jobs and resources
- Manage and monitor large-scale fleets of models in production

cnvrg simplifies ML workflows from end to end

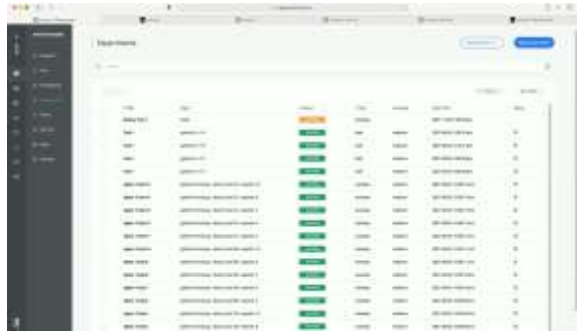
1 Create projects and workspaces



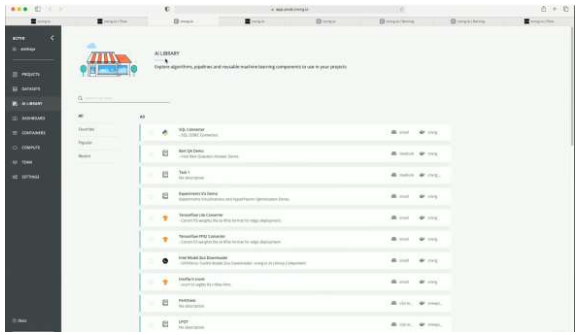
2 Connect data



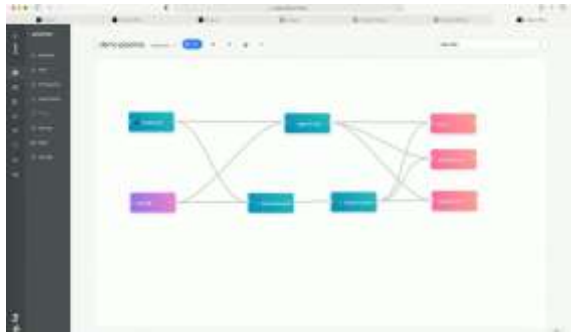
3 Manage experiments



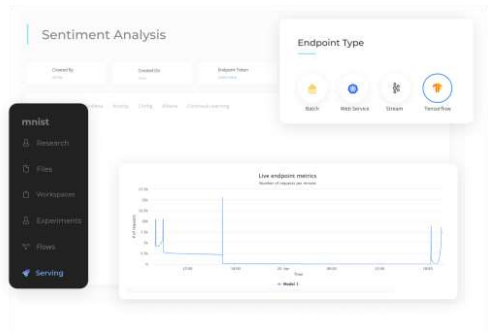
4 Create and re-use models



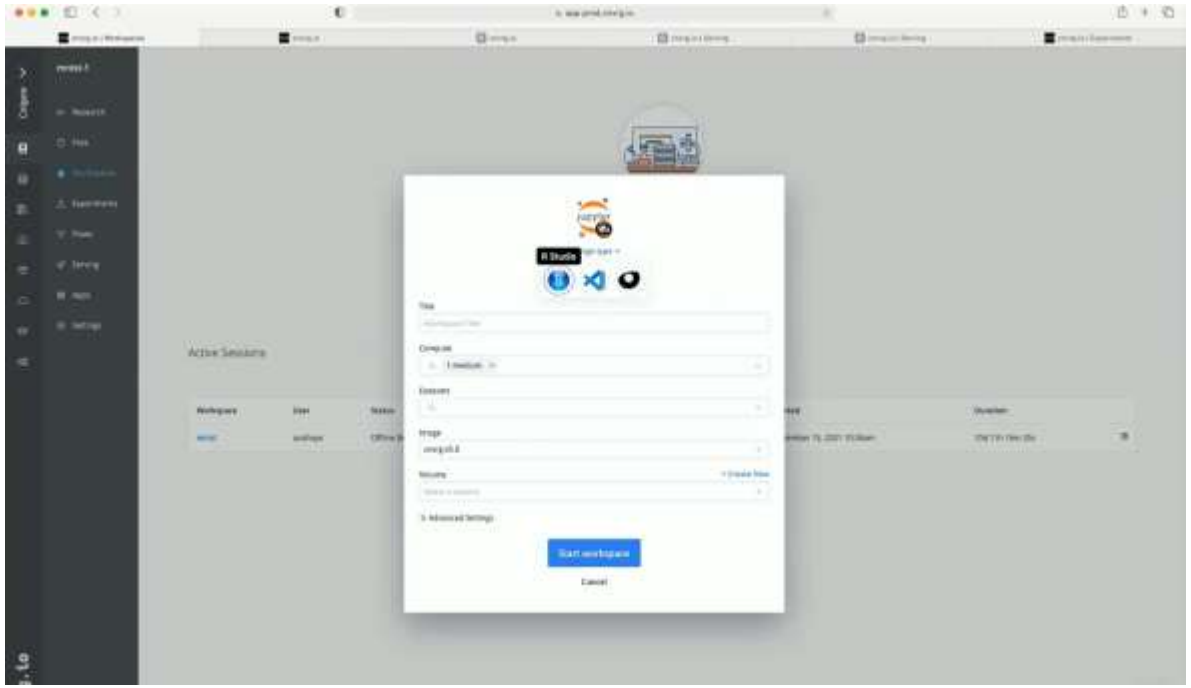
5 Drag-and-drop ML pipelines



6 Deploy and monitor models/clusters

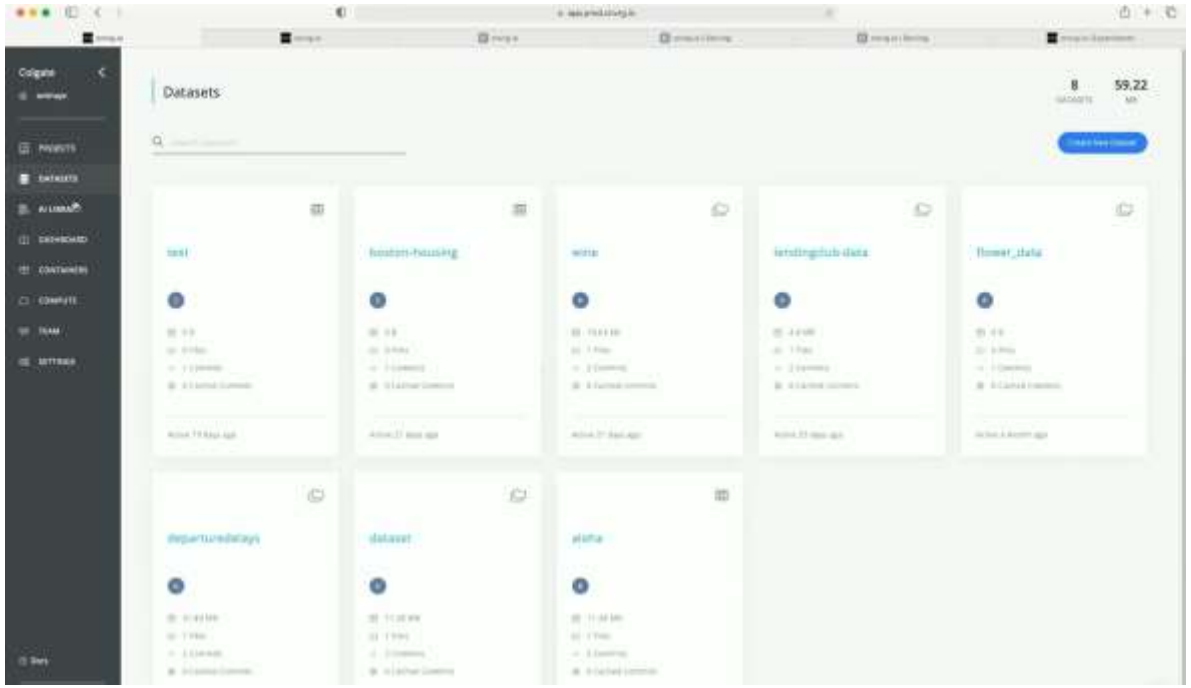


1 Start with projects and workspaces



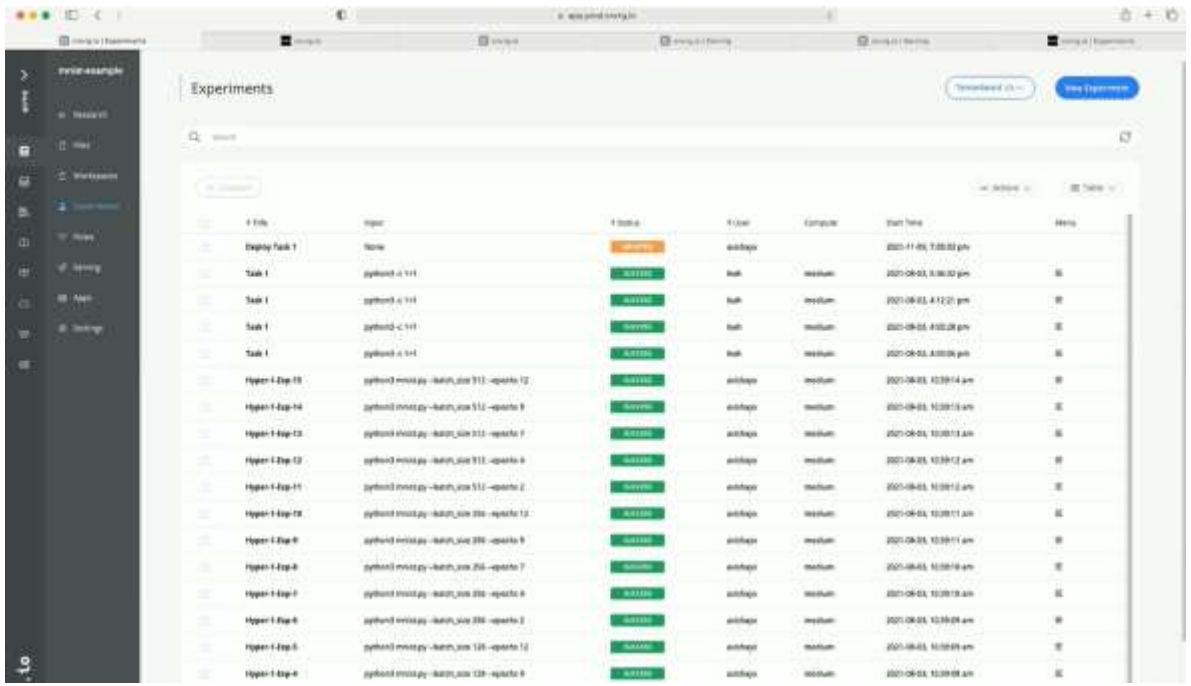
- Organize datasets, models, artifacts in one central place
- Explore and experiment with popular notebooks and IDEs
- Share projects across teams effortlessly
- RBAC and encryption controls for more secure collaboration

2 Connect and work with data



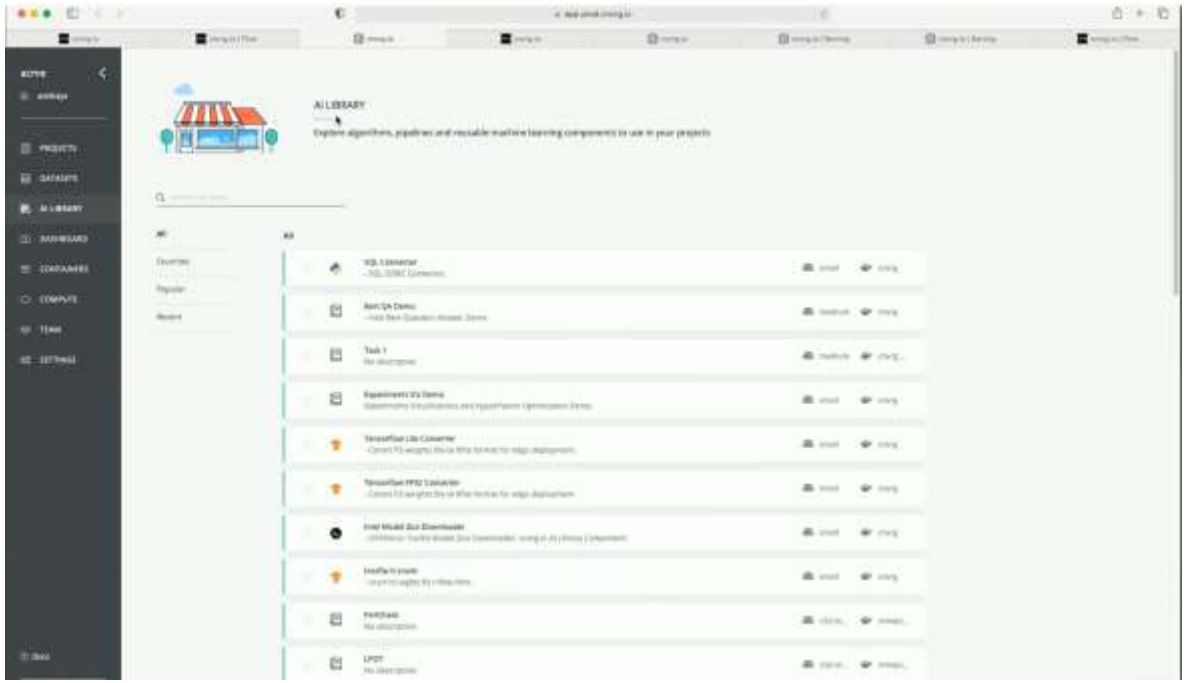
- Central data hub for all projects and users
- Connect to popular sources, types, formats
- Built-in versioning
- Built-in caching keeps data with compute, reduces the need for data movement

3 Run and manage experiments



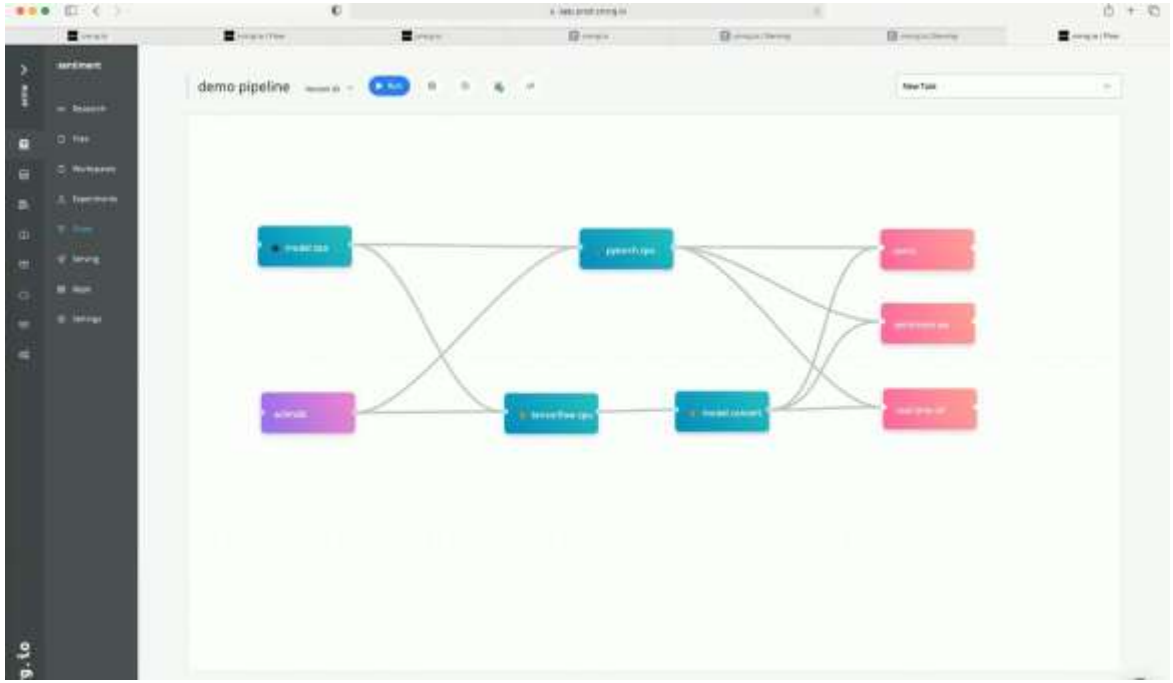
- Run thousands of experiments concurrently
- Users can track and compare models, hyperparameters, artifacts
- Supports Python, R, Java, Scala, other popular languages
- Run experiments on any local or cloud resource, or hybrid

4 Choose modular, reusable connectors, loaders, and algorithms



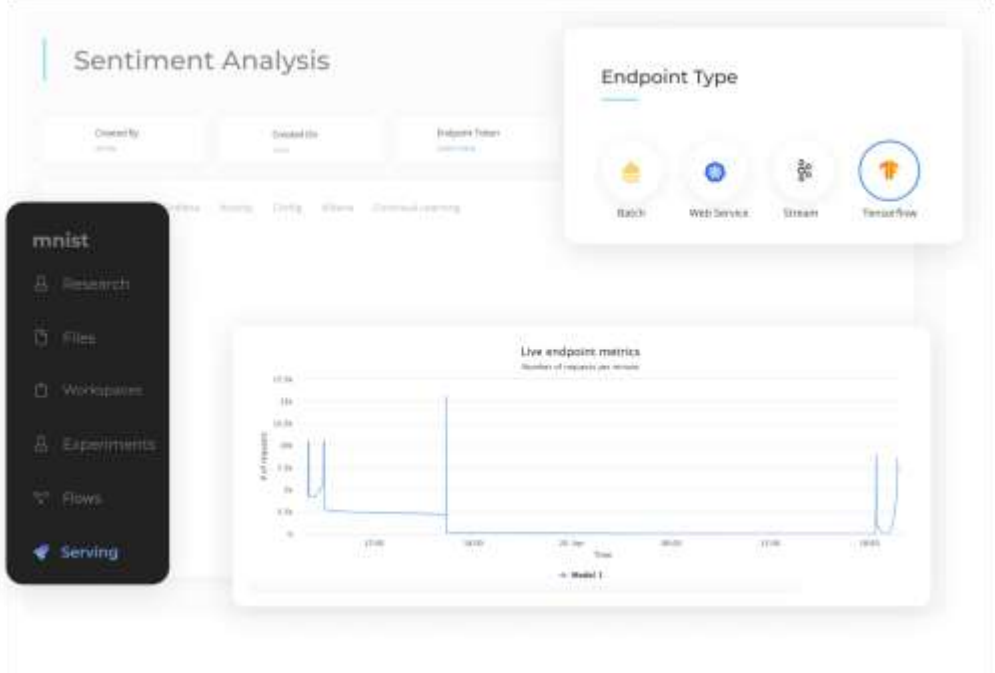
- Use your own or pre-packaged ML components from our AI Library
- Easy to customize, share, re-use
- Integration with Github, BitBucket to link your own model repos

5 Create drag-and-drop pipelines with Flows



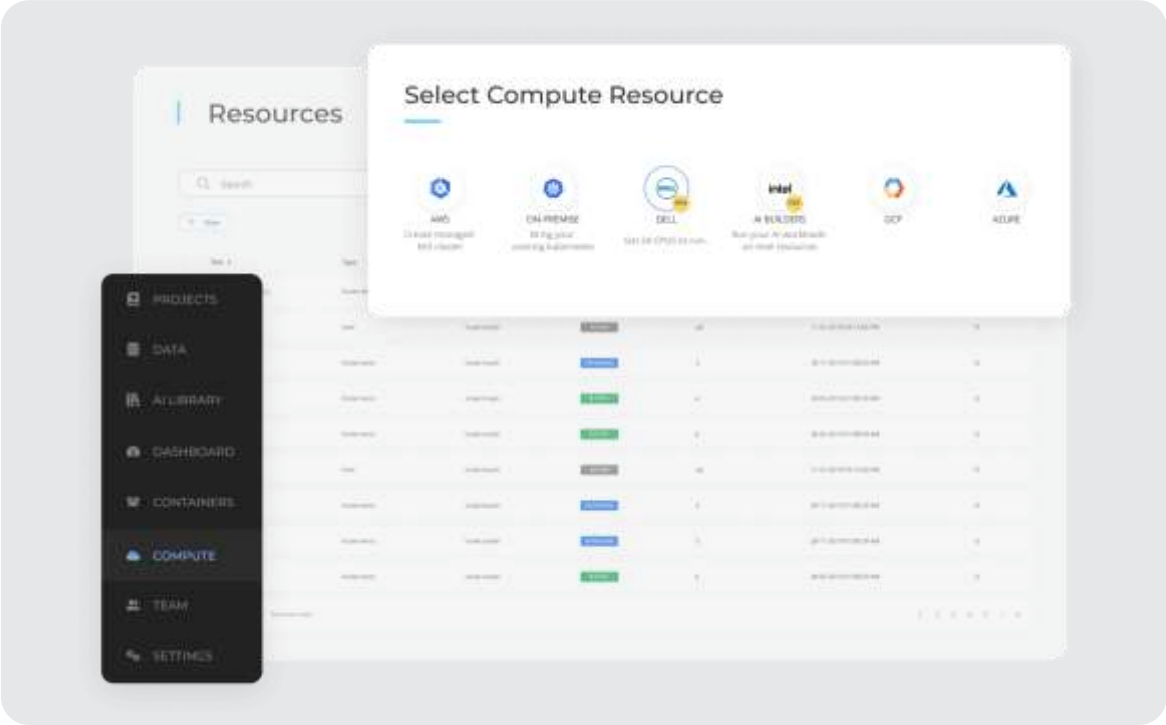
- Run tasks on your own hardware, or partner and OEM compute/storage from the cnvrg.io Marketplace
- Self-service CPU, GPU, tensor clusters, in a cloud-like way
- Run each pipeline stage on best-fit resource
- Auto-burst jobs from data center to clouds

6 Deploy and monitor models and clusters



- Deploy via web service, RabbitMQ, Kafka Streams
- Track model performance
- Trigger alerts or automate retraining
- Easy cluster and resource metrics with Grafana/Kibana dashboards

cnvrg.io Marketplace:
AI infrastructure with cloud-native simplicity

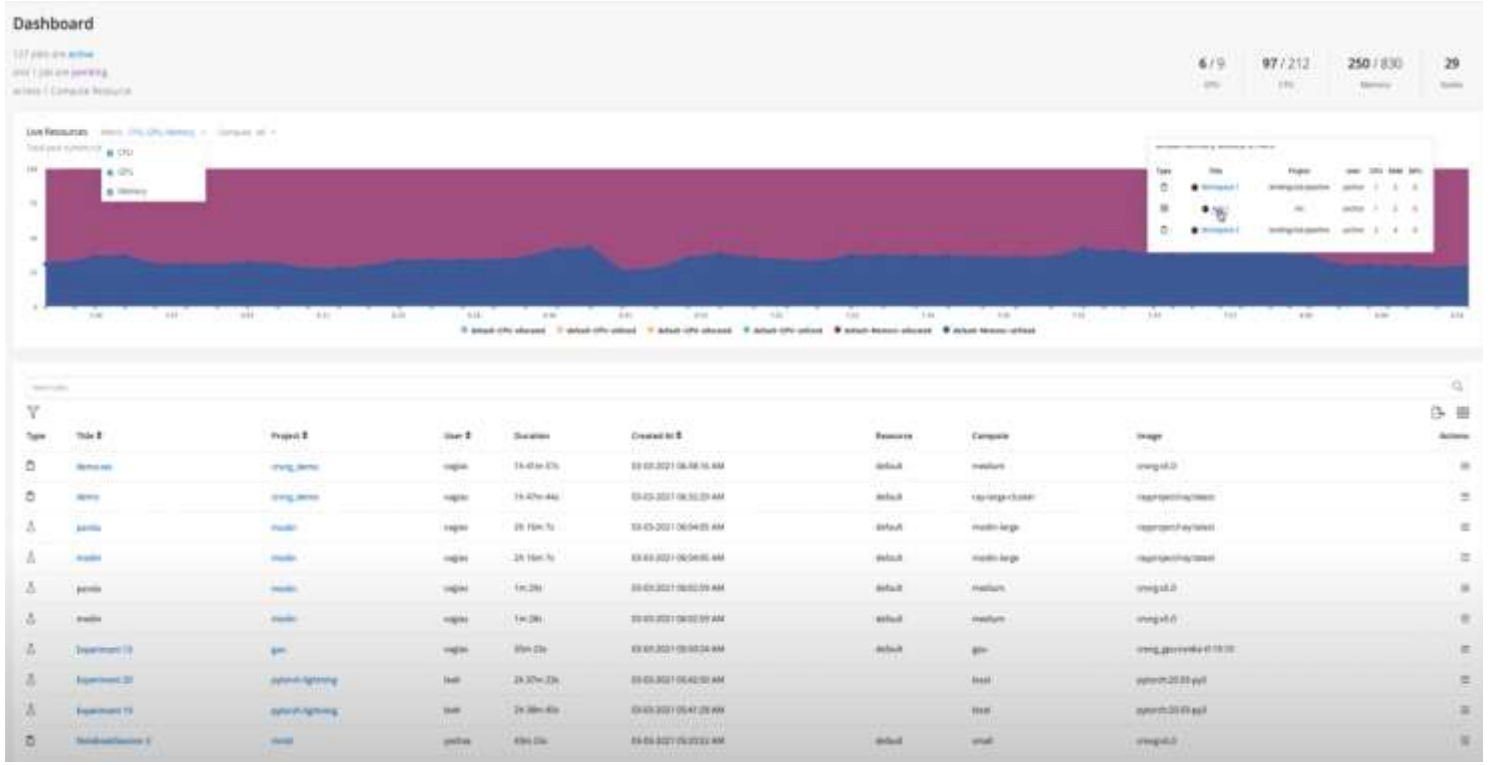


- Self-service selection of partner and OEM compute/storage
- Choose CPUs, GPUs, tensors, AI accelerators, storage in customizable sizes
- Choose partner clouds, or cloud-like deployment on-prem
- No more waiting for infra and ops teams to stand up resources
- Menu-driven, point and click configuration settings

Monitoring resource health and consumption is easy with cnvrg.io

From the dashboard, you can see CPU, GPU, and memory used and allocated

Granular visibility down to users and jobss



cnvrg.io works with all popular K8s distributions

All major CNCF-compliant K8s distributions, including

- OpenShift
- VMware
- EKS, AKS, GKS

cnvrg.io control plane contains all of the MLOps logic – installs as a K8s operator on any supported distribution

K8s taints and tolerations restrict pods to appropriate clusters, e.g., GPU jobs only running on GPU-enabled clusters

Smart Manufacturing: deploying defect detection with MLOps at scale

Challenge

- Siloed legacy architecture/ manual workflows
- Underutilization of hybrid cloud resources

Requirements

- Needed to deploy defect detection across global manufacturing facilities
- Improvement of data science efficiency and productivity

Results

- Improved data scientists' efficiency by 50% with MLOps automation
- Accelerated infrastructure transformation to modern AI workflow
- Successfully demonstrated scalable AI deployment across global facilities
- Collaboration globally across advanced analytics, engineering and IT teams
- Maximized on prem & hybrid cloud node utilization with scale to zero

Achieving Massive Business Growth with Large-Scale Models in Production

Wargaming.net

Challenge

- SAS offered little flexibility & scalability
- Underutilization of on premises & wanted to transition to cloud

Requirements

- Support and automation of more than 1500 models in production
- Ability to increase performance with cloud & accelerator adoption

Results

- Automated pipelines for 1500+ models in production
- Enabled seamless cloud adoption (AWS) + 100% hardware utilization
- Reduced cost by more than 50% on servers and licensing
- Accelerated time to production by 30% by eliminating bottlenecks

WARGAMING.NET

dmlc XGBoost

APACHE Spark

python

Dash by plotly

Get Started Today!

Metacloud
at cnvrg.io/blog/

Docs
app.cnvrg.io/docs/

Cnvrg.io on YouTube
tutorials, examples, and proven practices

Thank you

cnvrg.io