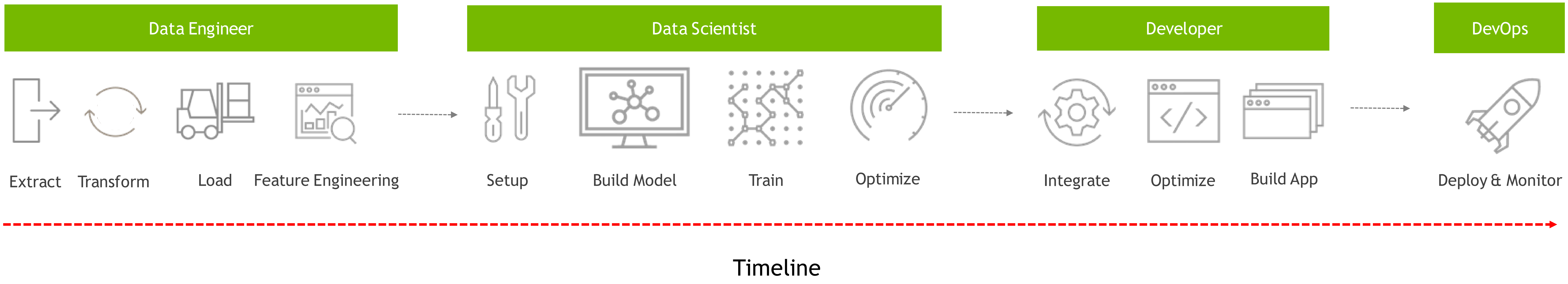
BUILD AI SOLUTIONS WITH NVIDIA NGC AND RED HAT OPENSHIFT

Abhishek Sawarkar, Chintan Patel. NVIDIA
Deepthi Dharwar, Diane Feddema. Red Hat

AI WORKFLOWS ARE COMPLEX



Data Engineer: Extract, Transform, Load, Feature Engineering

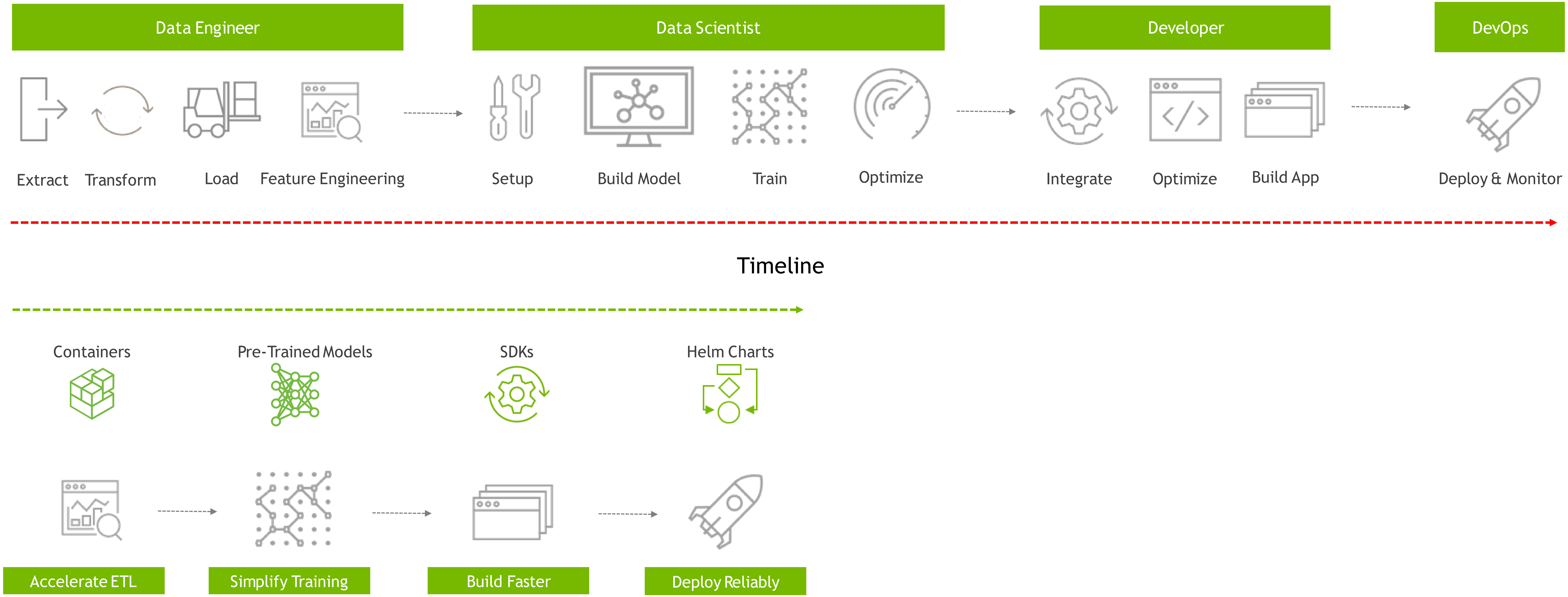
Data Scientist: Setup, Build Model, Train, Optimize

Developer: Integrate, Optimize, Build App

DevOps: Deploy & Monitor

Timeline

NGC CATALOG HELPS SIMPLIFY AND ACCELERATE AI WORKFLOWS



Data Engineer: Extract, Transform, Load, Feature Engineering

Data Scientist: Setup, Build Model, Train, Optimize

Developer: Integrate, Optimize, Build App

DevOps: Deploy & Monitor

Timeline

Containers

Pre-Trained Models

SDKs

Helm Charts

Accelerate ETL

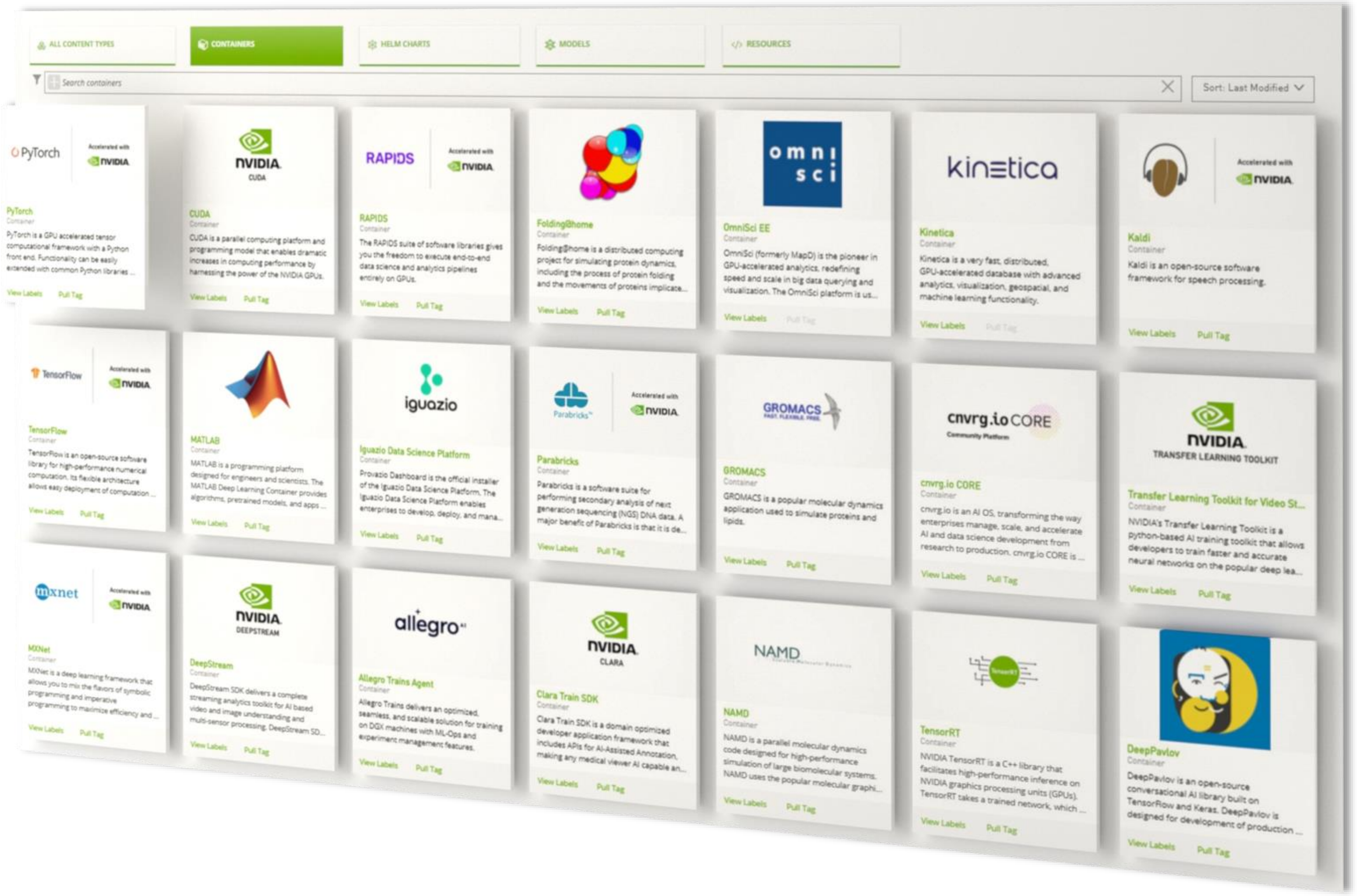
Simplify Training

Build Faster

Deploy Reliably

NVIDIA. Red Hat

NGC CONTAINERS ENABLE YOU TO FOCUS ON BUILDING AI



PyTorch
Container
PyTorch is a GPU accelerated tensor computational framework with a Python front end. Functionality can be easily extended with common Python libraries ...

CUDA
Container
CUDA is a parallel computing platform and programming model that enables dramatic increases in computing performance by harnessing the power of the NVIDIA GPUs.

RAPIDS
Container
The RAPIDS suite of software libraries gives you the freedom to execute end-to-end data science and analytics pipelines entirely on GPUs.

Folding@home
Container
Folding@home is a distributed computing project for simulating protein dynamics, including the process of protein folding and the movements of proteins implicate...

OmniSci EE
Container
OmniSci (formerly MapD) is the pioneer in GPU-accelerated analytics, redefining speed and scale in big data querying and visualization. The OmniSci platform is us...

Kinetica
Container
Kinetica is a very fast, distributed, GPU-accelerated database with advanced analytics, visualization, geospatial, and machine learning functionality.

Kaldi
Container
Kaldi is an open-source software framework for speech processing.

TensorFlow
Container
TensorFlow is an open-source software library for high-performance numerical computation. Its flexible architecture allows easy deployment of computation ...

MATLAB
Container
MATLAB is a programming platform designed for engineers and scientists. The MATLAB Deep Learning Container provides algorithms, pretrained models, and apps ...

Iguazio Data Science Platform
Container
Provazio Dashboard is the official installer of the Iguazio Data Science Platform. The Iguazio Data Science Platform enables enterprises to develop, deploy, and mana...

Parabricks
Container
Parabricks is a software suite for performing secondary analysis of next generation sequencing (NGS) DNA data. A major benefic of Parabricks is that it is de...

GROMACS
Container
GROMACS is a popular molecular dynamics application used to simulate proteins and lipids.

cnvrg.io CORE
Container
cnvrg.io is an AI OS, transforming the way enterprises manage, scale, and accelerate AI and data science development from research to production. cnvrg.io CORE is ...

Transfer Learning Toolkit for Video St...
Container
NVIDIA's Transfer Learning Toolkit is a python-based AI training toolkit that allows developers to train faster and accurate neural networks on the popular deep lea...

MXNet
Container
MXNet is a deep learning framework that allows you to mix the flavors of symbolic programming and imperative programming to maximize efficiency and ...

DeepStream
Container
DeepStream SDK delivers a complete streaming analytics toolkit for AI based video and image understanding and multi-sensor processing. DeepStream SD...

Allegro Trains Agent
Container
Allegro Trains delivers an optimized, seamless, and scalable solution for training on DGX machines with ML-Ops and experiment management features.

Clara Train SDK
Container
Clara Train SDK is a domain optimized developer application framework that includes APIs for AI-Assisted Annotation, making any medical viewer AI capable an...

NAMD
Container
NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. NAMD uses the popular molecular graphi...

TensorRT
Container
NVIDIA TensorRT is a C++ library that facilitates high-performance inference on NVIDIA graphics processing units (GPUs). TensorRT takes a trained network, which ...

DeepPavlov
Container
DeepPavlov is an open-source conversational AI library built on TensorFlow and Keras. DeepPavlov is designed for development of production ...

ENTERPRISE READY SOFTWARE

Scanned for CVEs, malware, crypto

Tested for reliability

Backed by Enterprise support

PERFORMANCE OPTIMIZED

Scalable

Updated Monthly

Better performance on the same system

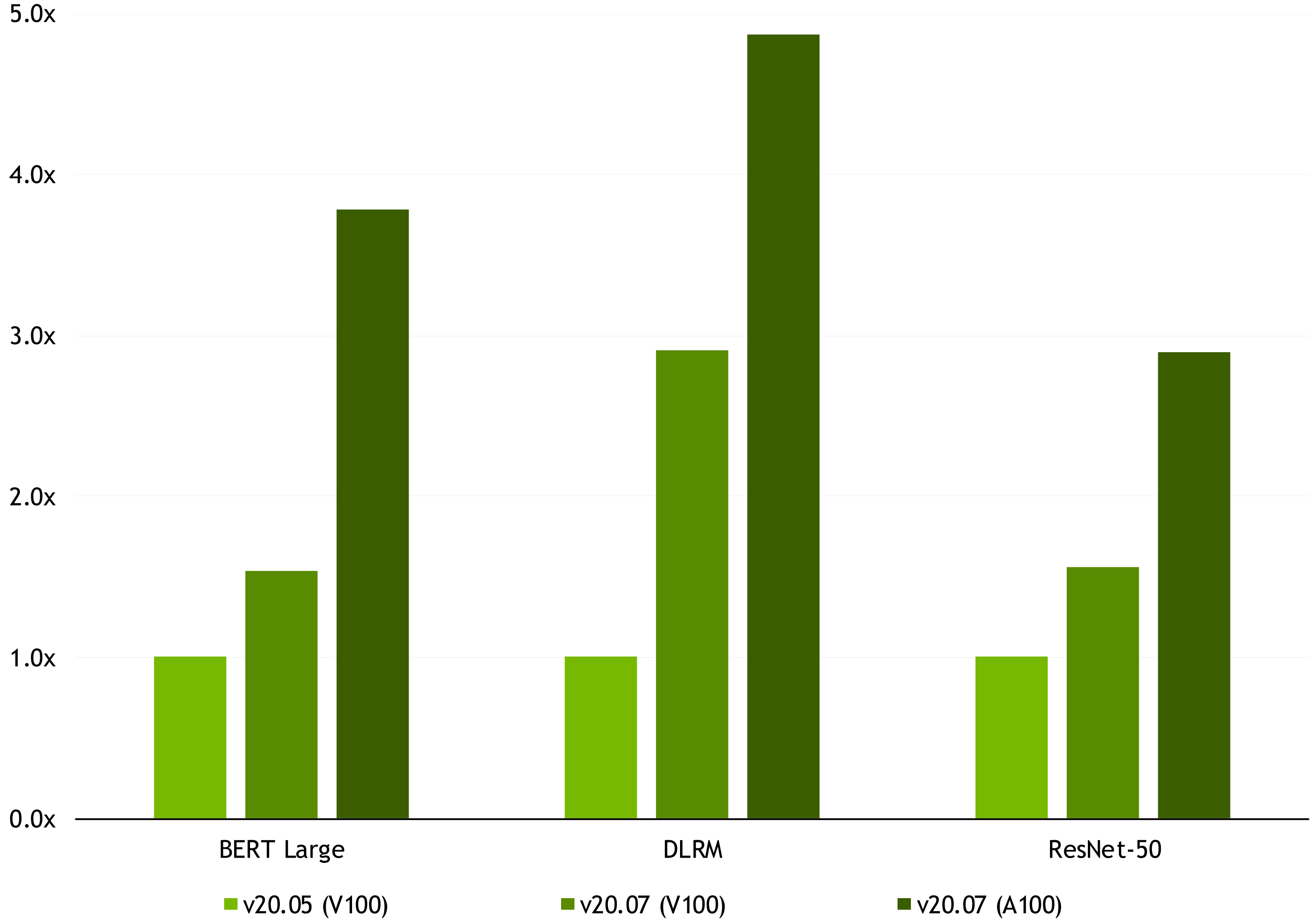
DEPLOY ANYWHERE

Docker, cri-o runtimes

Bare metal, VMs, Kubernetes

Multi-cloud, on-prem, hybrid, edge

DO WHAT YOU DO BEST, FASTER



ENTERPRISE READY SOFTWARE

Scanned for CVEs, malware, crypto

Tested for reliability

Backed by Enterprise support

PERFORMANCE OPTIMIZED

Scalable

Updated Monthly

Better performance on the same system

DEPLOY ANYWHERE

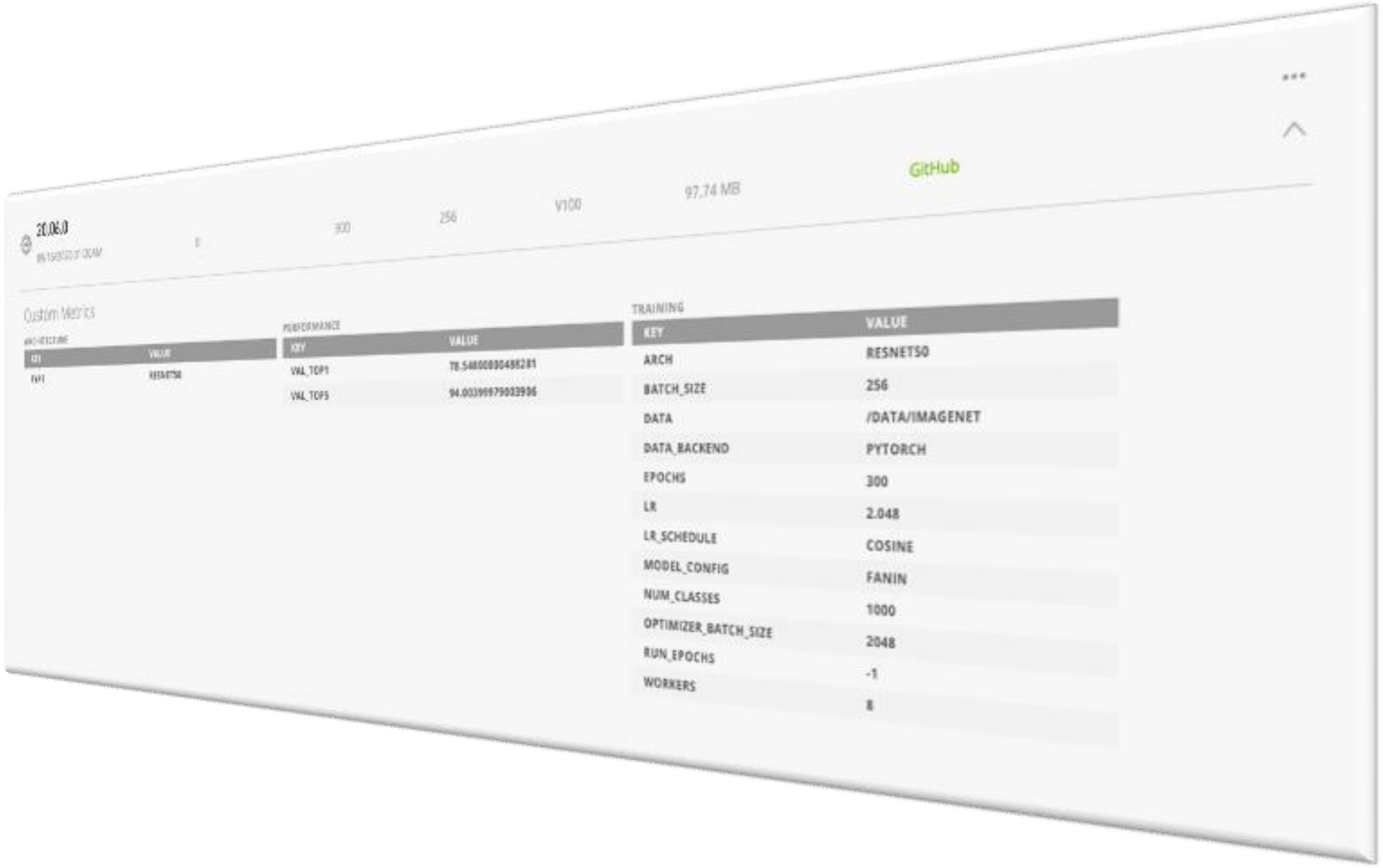
Docker, Singularity runtimes

Bare metal, VMs, Kubernetes

Multi-cloud, on-prem, hybrid, edge

BERT-Large and ResNet-50 v1.5 Training performance with TensorFlow on a single node 8x V100 (32GB) & A100 (40GB). Mixed Precision. Batch size for BERT: 10 (V100), 24 (A100), ResNet: 512 (V100, v20.05), 256 (v20.07)
DLRM Training performance with PyTorch on 1x V100 & 1x A100. Mixed Precision. Batch size 32768. DRLM trained with v20.03 and v20.07

EASILY IDENTIFY THE RIGHT MODELS WITH CREDENTIALS



20.06.0 [illegible] 8 300 256 V100 97.74 MB GitHub

Custom Metrics

ARCHITECTURE

KEY	VALUE
[illegible]	RESNET50

PERFORMANCE

KEY	VALUE
VAL_TOP1	78.54800800488281
VAL_TOP5	94.00399879003906

TRAINING

KEY	VALUE
ARCH	RESNET50
BATCH_SIZE	256
DATA	/DATA/IMAGENET
DATA_BACKEND	PYTORCH
EPOCHS	300
LR	2.048
LR_SCHEDULE	COSINE
MODEL_CONFIG	FANIN
NUM_CLASSES	1000
OPTIMIZER_BATCH_SIZE	2048
RUN_EPOCHS	-1
WORKERS	8

WIDE RANGE OF USE CASES

ResNet-50, SSD, MobileNet, VGG16

WaveGlow, BERT, NeMo

Wide & Deep, DLRM & many more

PRE-TRAINED MODELS

Faster training

Higher accuracy

Transparency through credentials

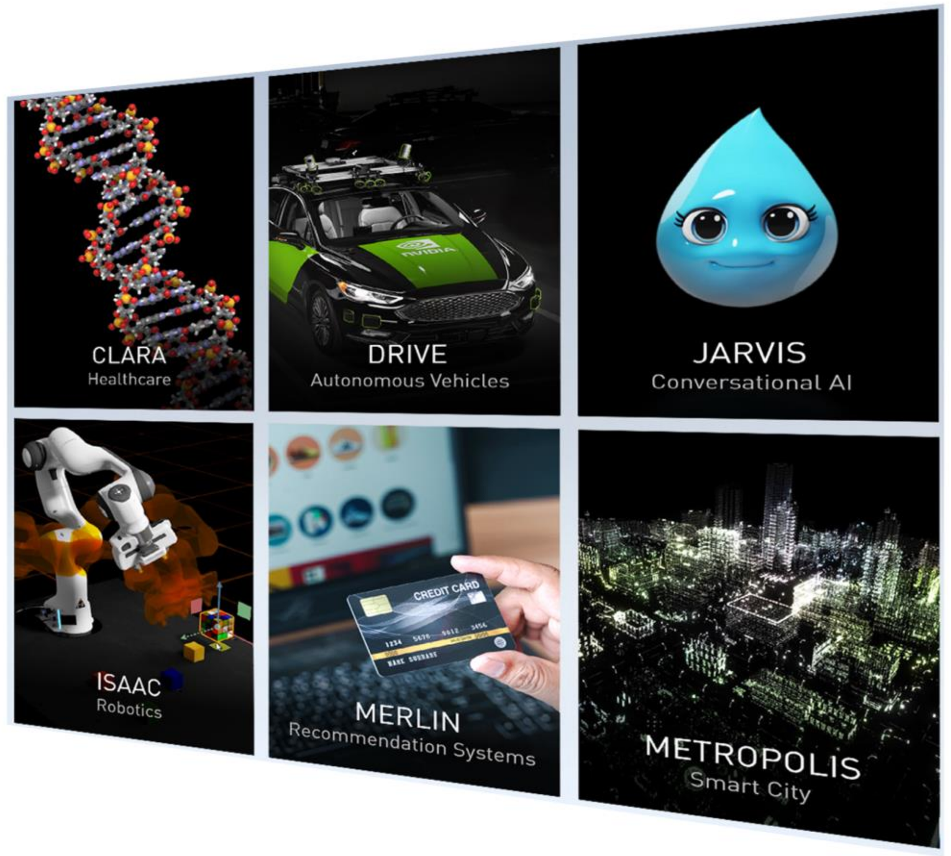
RESOURCES

Get started with code samples

Customize NGC models

Reproduce with recipes

INDUSTRY APP FRAMEWORKS FOR END-TO-END AI WORKFLOWS



TRANSFER LEARNING TOOLKIT

Domain adaptability

Significantly reduce development time

TENSORRT

Optimizes for low latency and high-throughput

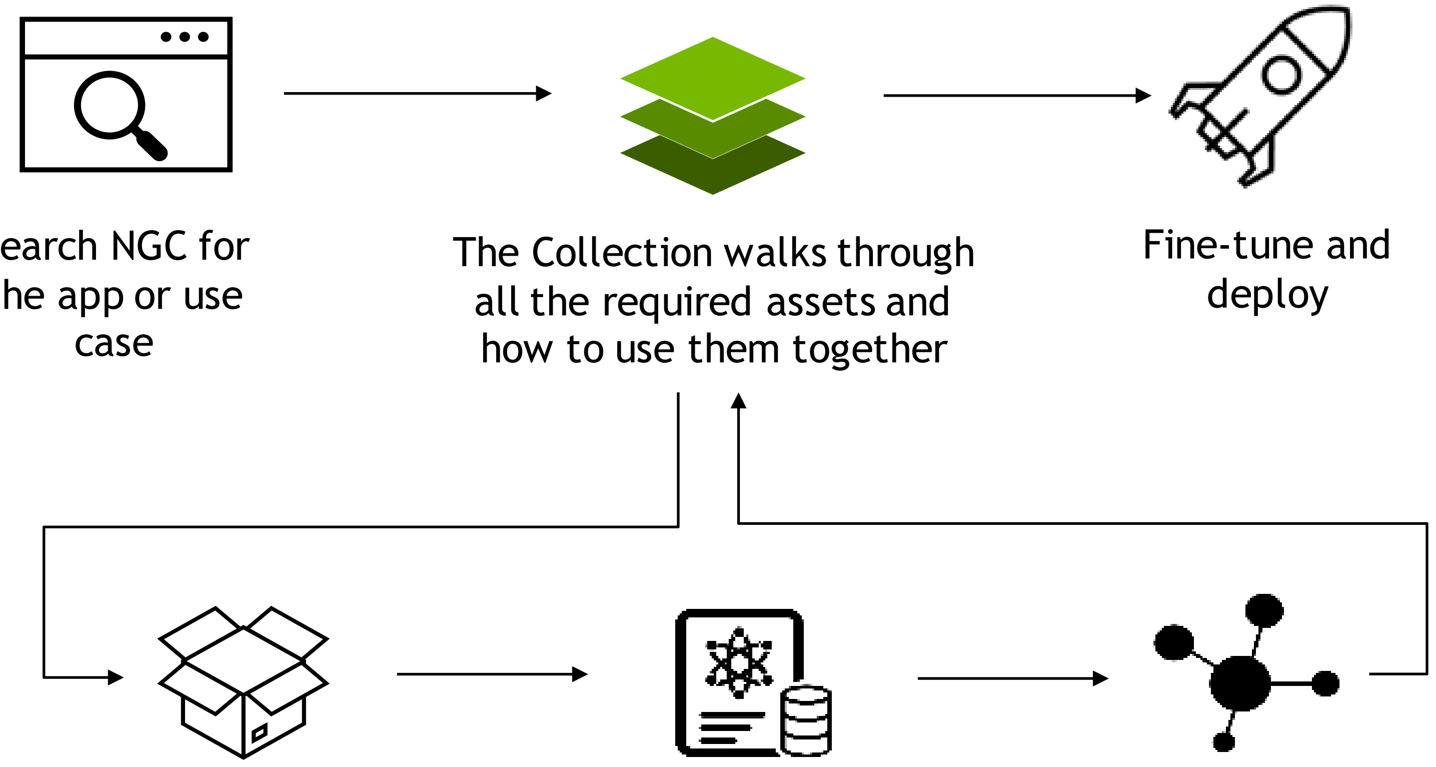
Integrated with major frameworks

TRITON

High performance inference on GPU/CPU systems

Supports multiple frameworks backends

EVERYTHING YOU NEED TO BUILD YOUR AI IN ONE LOCATION



Search NGC for the app or use case

The Collection walks through all the required assets and how to use them together

Fine-tune and deploy

Deploy container

Start Jupyter Notebook instance

Download model

COLLECTIONS

Compatible assets grouped together, removes guesswork

Curated software by use cases

Detailed documentation further simplifies work for users

READY-TO-USE

Conversational AI

Computer Vision

NVIDIA AI App Frameworks

WHY KUBERNETES AND DEVOPS FOR AI/ML?

AGILITY

Respond quickly
with automated compute resource
management, and increased
collaboration

CONSISTENCY & PORTABILITY

Develop and deploy ML models consistently across data center, edge, and public clouds.

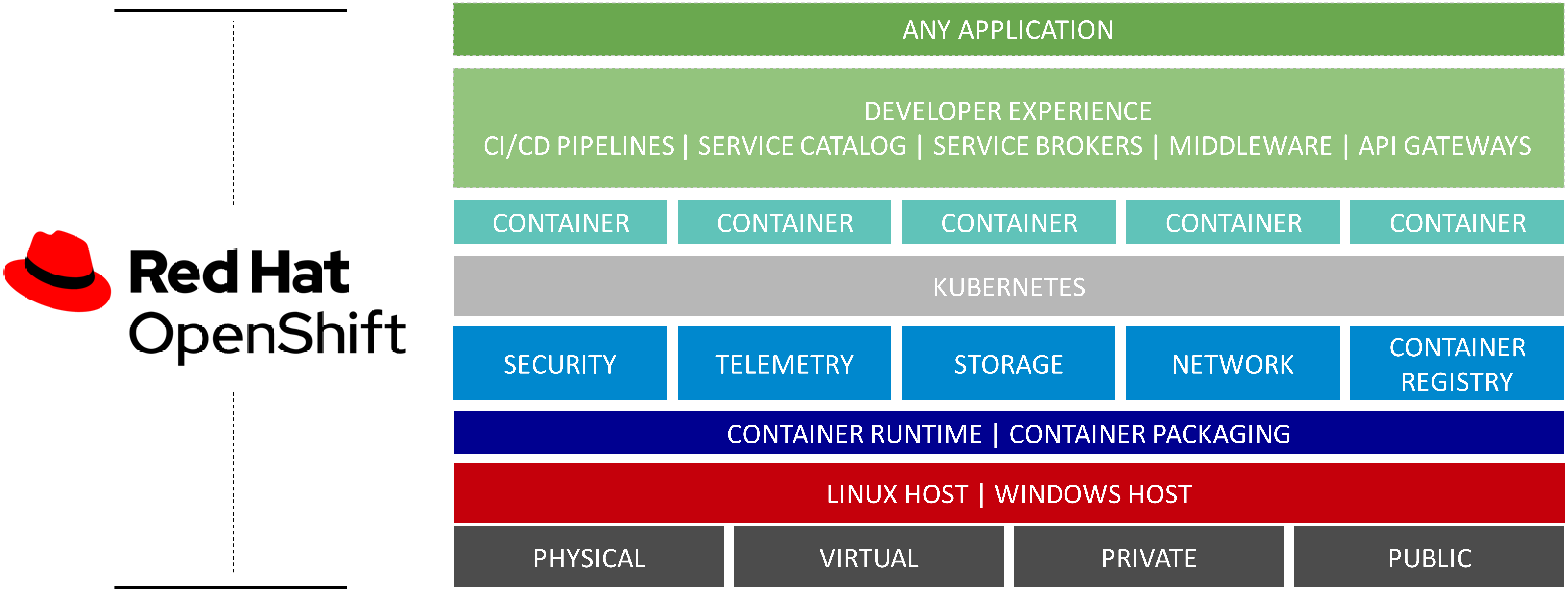
FLEXIBILITY

Provision AI/ML environments as and when you need them.

SCABILITY

Autoscaling and high availability of the AI/ML solution stack.

OPENSHIFT ENABLES CONTAINERS, KUBERNETES, AND DEVOPS IN PRODUCTION



Red Hat OpenShift

ANY APPLICATION

DEVELOPER EXPERIENCE
CI/CD PIPELINES | SERVICE CATALOG | SERVICE BROKERS | MIDDLEWARE | API GATEWAYS

CONTAINER | CONTAINER | CONTAINER | CONTAINER | CONTAINER

KUBERNETES

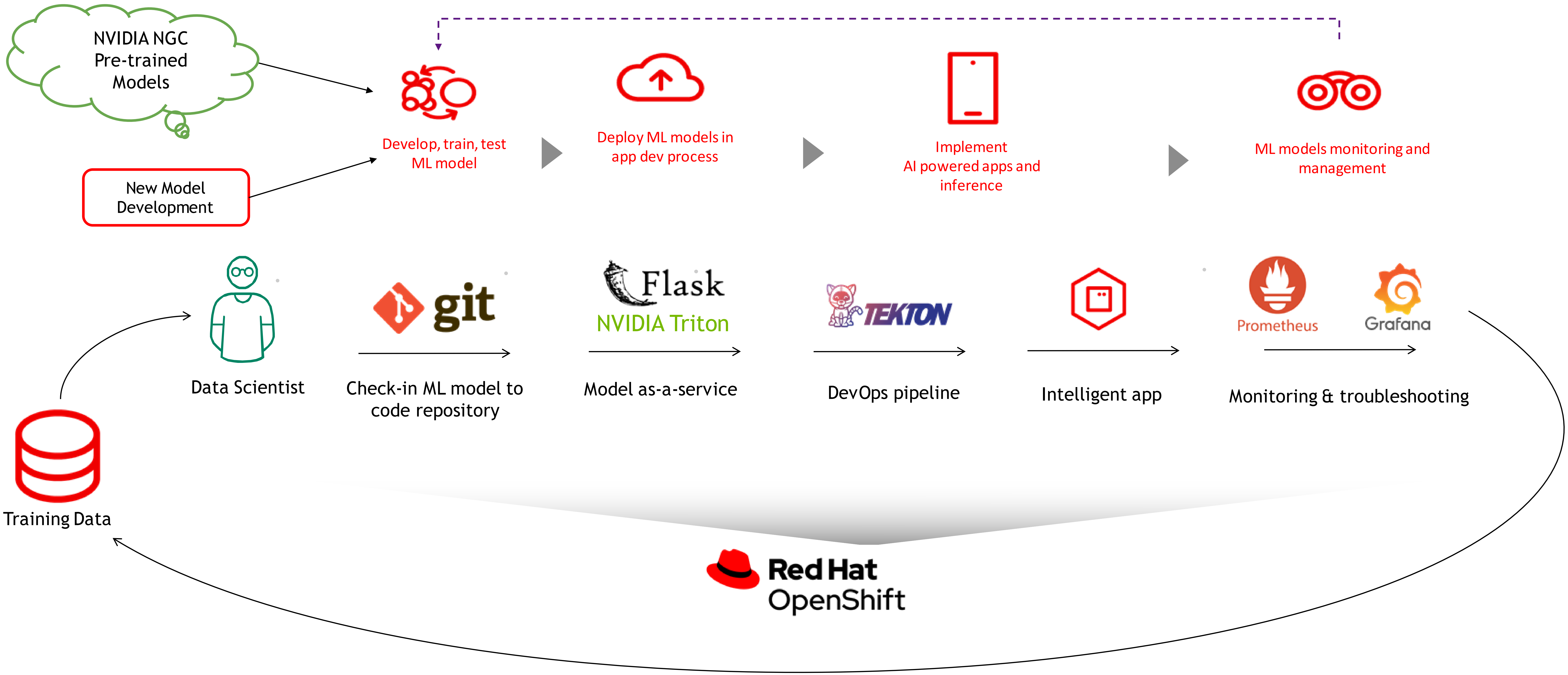
SECURITY | TELEMETRY | STORAGE | NETWORK | CONTAINER REGISTRY

CONTAINER RUNTIME | CONTAINER PACKAGING

LINUX HOST | WINDOWS HOST

PHYSICAL | VIRTUAL | PRIVATE | PUBLIC

RED HAT OPENSHIFT HELPS FAST TRACK AI/ML LIFECYCLE



NVIDIA NGC Pre-trained Models

New Model Development

Develop, train, test ML model

Deploy ML models in app dev process

Implement AI powered apps and inference

ML models monitoring and management

Data Scientist

git

Check-in ML model to code repository

Flask

NVIDIA Triton

Model as-a-service

TEKTON

DevOps pipeline

Intelligent app

Prometheus

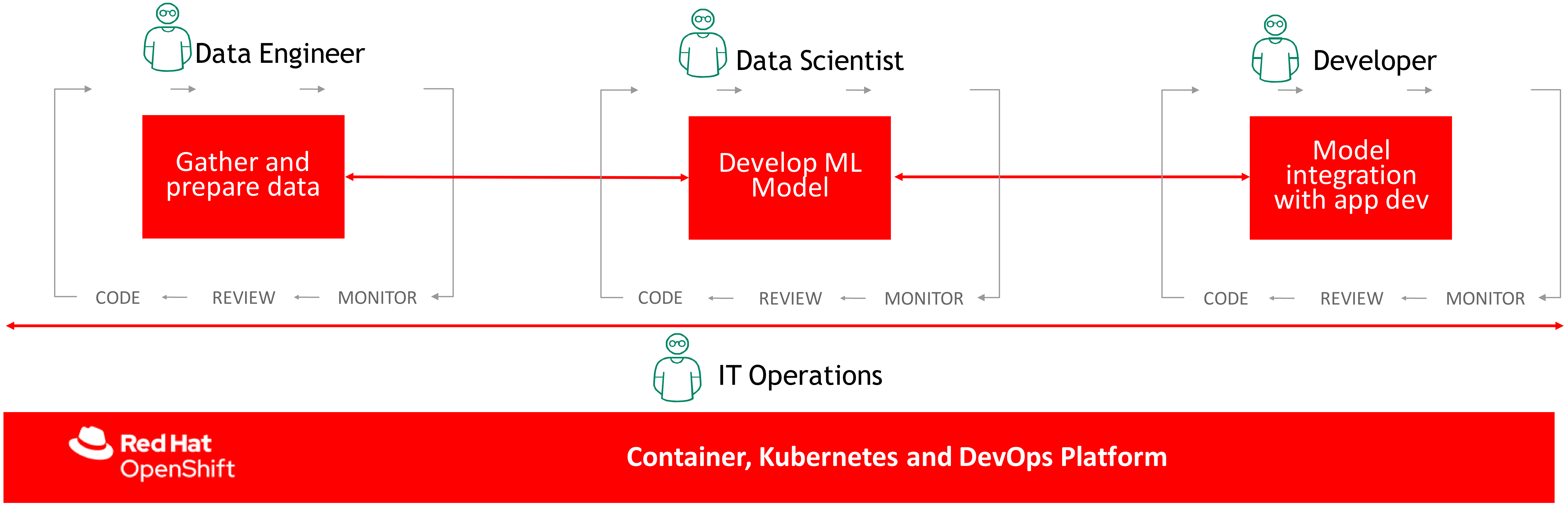
Grafana

Monitoring & troubleshooting

Training Data

Red Hat OpenShift

RED HAT OPENSHIFT HELPS FAST TRACK AI/ML LIFECYCLE



Data Engineer

Gather and prepare data

CODE ← REVIEW ← MONITOR

Data Scientist

Develop ML Model

CODE ← REVIEW ← MONITOR

Developer

Model integration with app dev

CODE ← REVIEW ← MONITOR

IT Operations

Red Hat OpenShift

Container, Kubernetes and DevOps Platform

- File (NFS, HDFS), Object (S3) and Block
- High Throughput, Low Latency, Secure
- Data Movement - Kafka
- Data Analytics - Spark and BDC
- Data pipelines - Tekton and ArgoCD

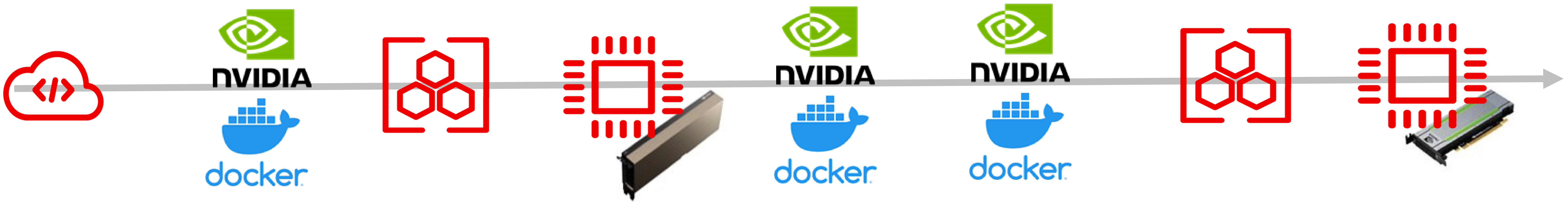
- CPUs, memory, GPUs, FPGA
- High speed networking and storage
- Containers with - Language (python), frameworks (PyTorch), IDE (Jupyter)
- On-demand scale up
- DevOps - Build v2, Tekton, ArgoCD

- RESTful services for models
- Services monitoring and alerting
- Services logging and diagnostics

CONVERSATIONAL AI DEMO

TASK: TRAIN, OPTIMIZE & DEPLOY A FINE-TUNED BERT MODEL

Demo Workflow



Training

Deployment

1. Load training data on OpenShift Container Storage

2. Get BERT container from NGC

3. Customize training yaml and mount persistent volume

4. Train on OpenShift using A100 or V100 or T4 GPUs

5. Optimize model using TensorRT

6. Get Triton container from NGC

7. Load on OpenShift Container Storage

8. Deploy on OpenShift using A100/V100 /T4 GPUs

Red Hat OpenShift

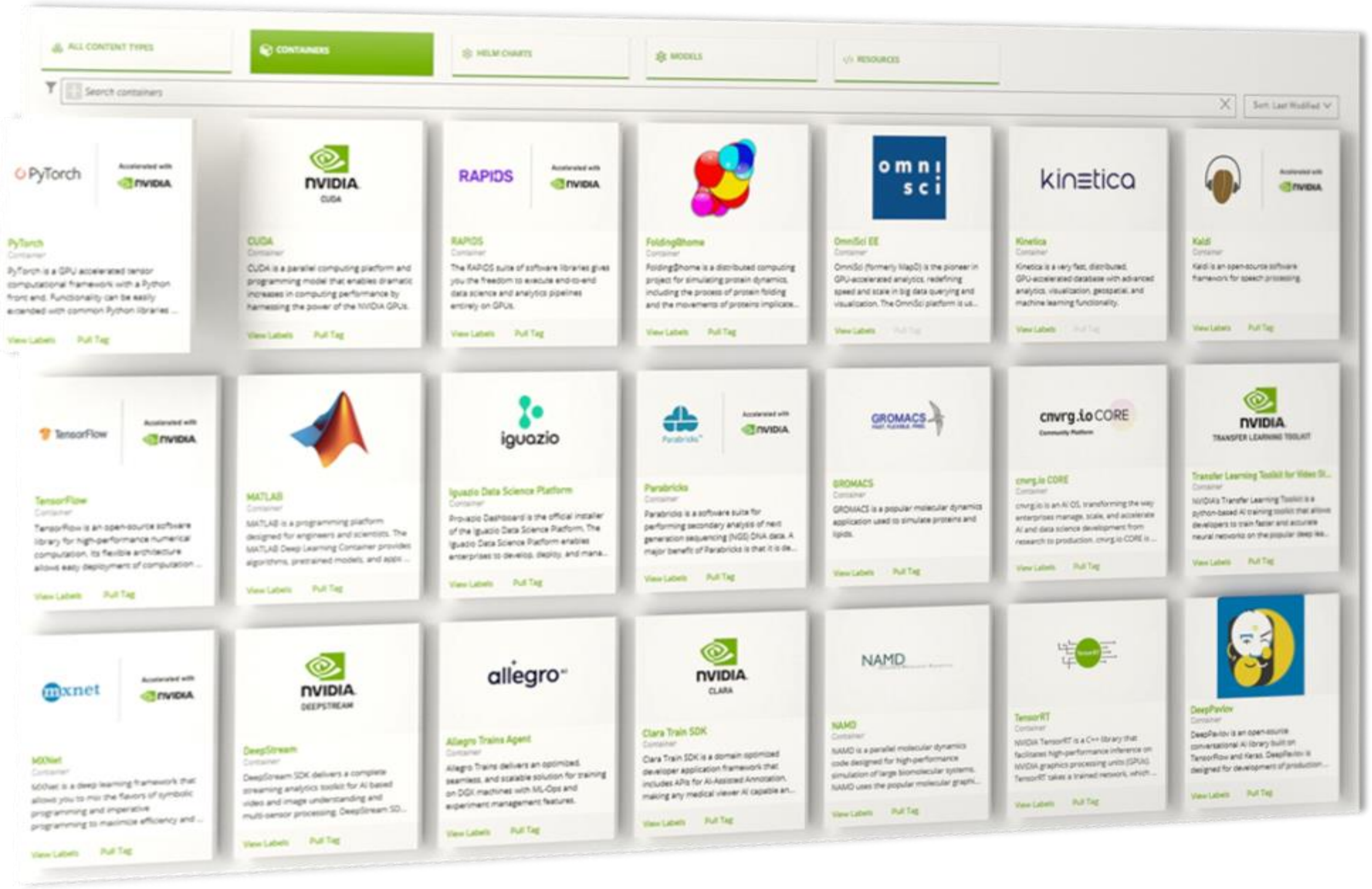
Red Hat OpenShift Container Storage

LET’S SEE IT IN ACTION

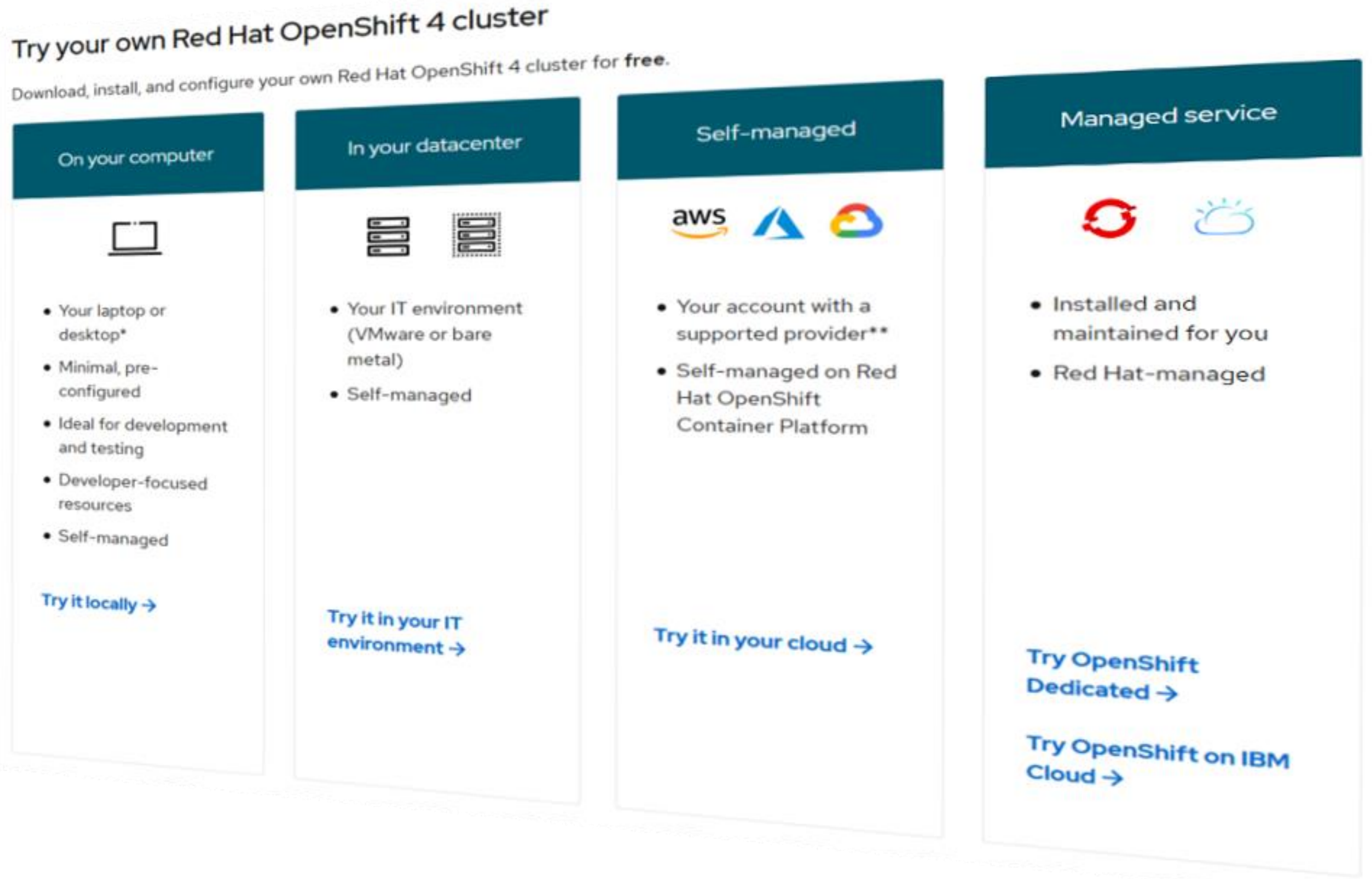
BUILD AI FASTER WITH NGC AND OPENSHIFT

Pull the OpenShift Collection from NGC and Run on OpenShift

ngc.nvidia.com



openshift.com/try | openshift.com/nvidia



Try your own Red Hat OpenShift 4 cluster

Download, install, and configure your own Red Hat OpenShift 4 cluster for **free**.

On your computer

- Your laptop or desktop*
- Minimal, pre-configured
- Ideal for development and testing
- Developer-focused resources
- Self-managed

Try it locally →

In your datacenter

- Your IT environment (VMware or bare metal)
- Self-managed

Try it in your IT environment →

Self-managed

- Your account with a supported provider**
- Self-managed on Red Hat OpenShift Container Platform

Try it in your cloud →

Managed service

- Installed and maintained for you
- Red Hat-managed

Try OpenShift Dedicated →

Try OpenShift on IBM Cloud →