intel® innovation

Intel® AI & Red Hat Solution

Agenda

- Intel® AI Portfolio Neil Dey (Intel)

- Red Hat and Intel AI Michael St Jean (Red Hat)
  - What is Red Hat® OpenShift Data Science (RHODS)
  - RHODS and Intel AI Portfolio
  - Demo: RHODS with DL1, AI Toolkit, OpenVINO™ Toolkit

- On-Prem AI Solution on OpenShift with Habana, AI Kit, OpenVINO, and cnvrg Neil Dey (Intel)
  - cnvrg.io overview (Bob Glithero - cnvrg)
  - Demo: OpenShift + cnvrg + AI Kit + OpenVINO + Gaudi - Blog (Bob Glithero - cnvrg)

The Habana® Gaudi® AI Training Processor

Designed to optimize AI performance, delivering higher AI efficiency than traditional CPUs and GPUs

Heterogeneous compute architecture enables high-efficiency on large AI workloads

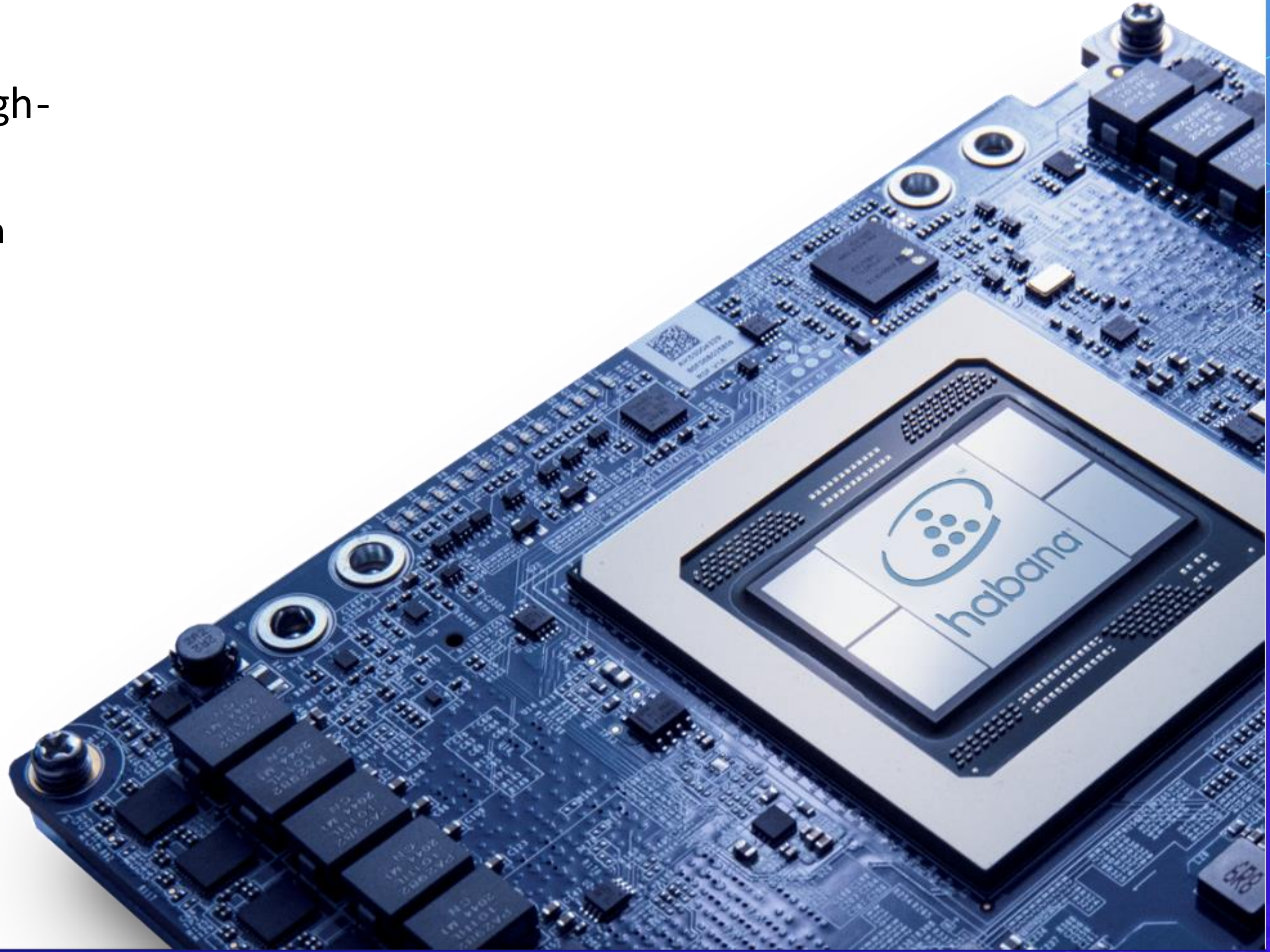
- GEMM engine (MME) excels at matrix multiplication
- While TPC runs non-linear and element wise ops

Software-managed memory architecture

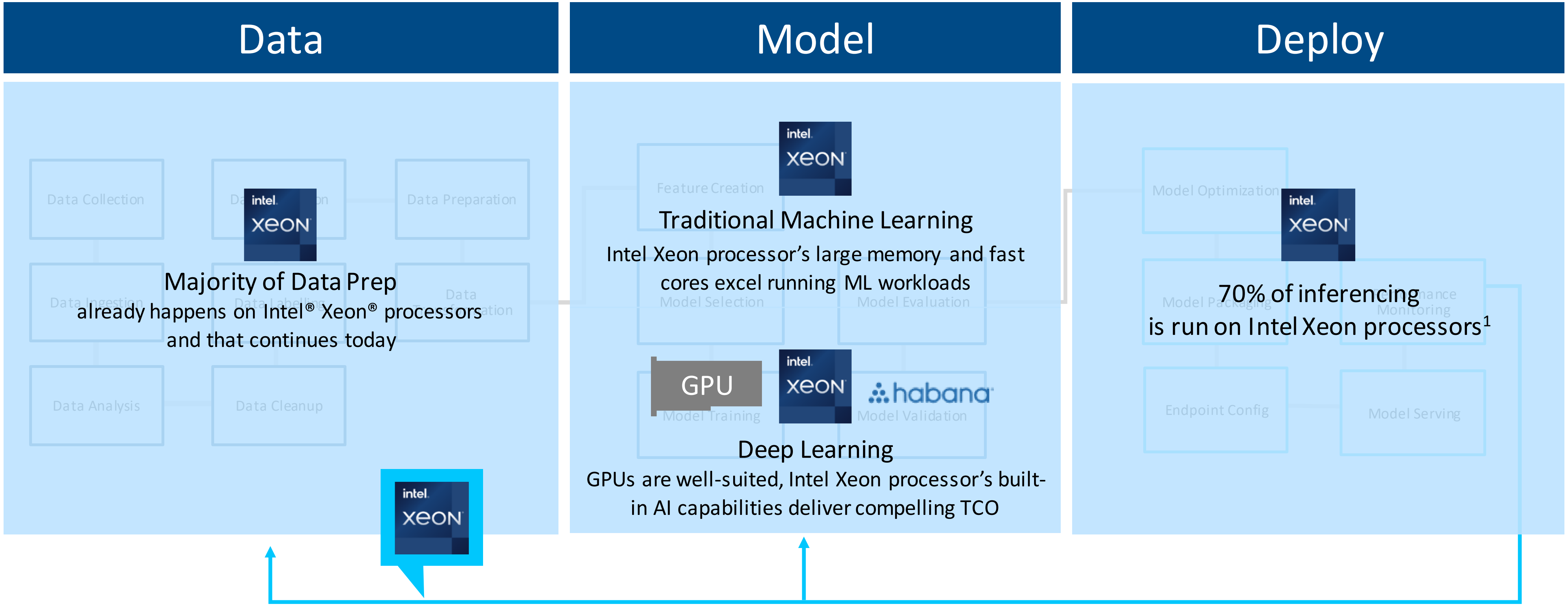
- 32 GB of HBM2 memory

Integrates ten 100Gb Ethernet RoCE ports

- Scaling capacity
- Flexibility based on industry standard
- Cost-efficiency with integrated NIC



The AI Pipeline Runs on Intel



Data

Majority of Data Prep already happens on Intel® Xeon® processors and that continues today

Model

Traditional Machine Learning

Intel Xeon processor’s large memory and fast cores excel running ML workloads

Deep Learning

GPUs are well-suited, Intel Xeon processor’s built-in AI capabilities deliver compelling TCO

Deploy

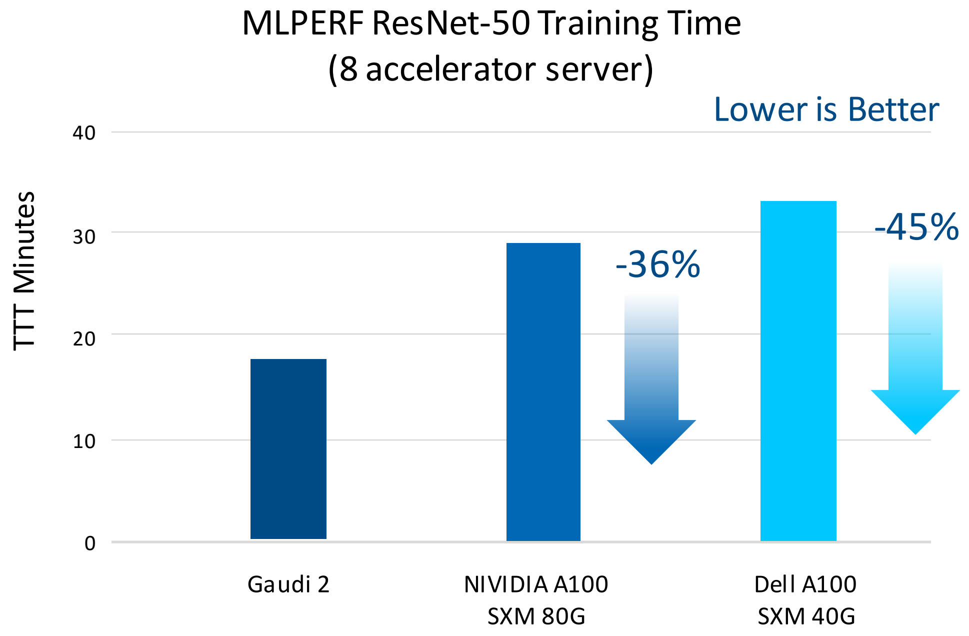
70% of inferencing is run on Intel Xeon processors[1]

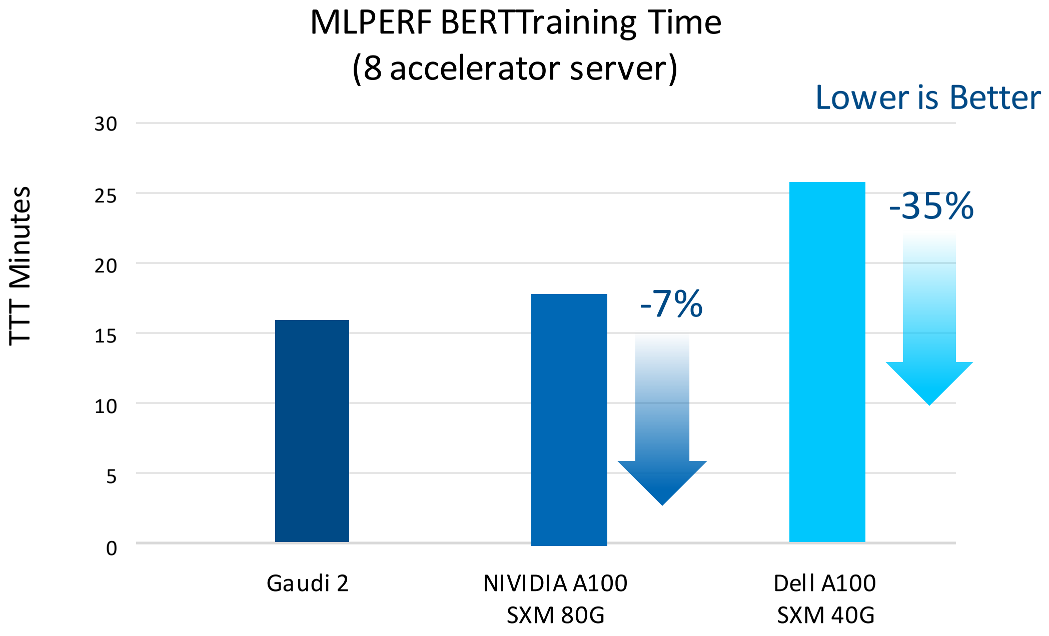
Database, Data Warehouse, Data Lake, Streaming Data, Feature Store, Model Registry

[1] Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2021.

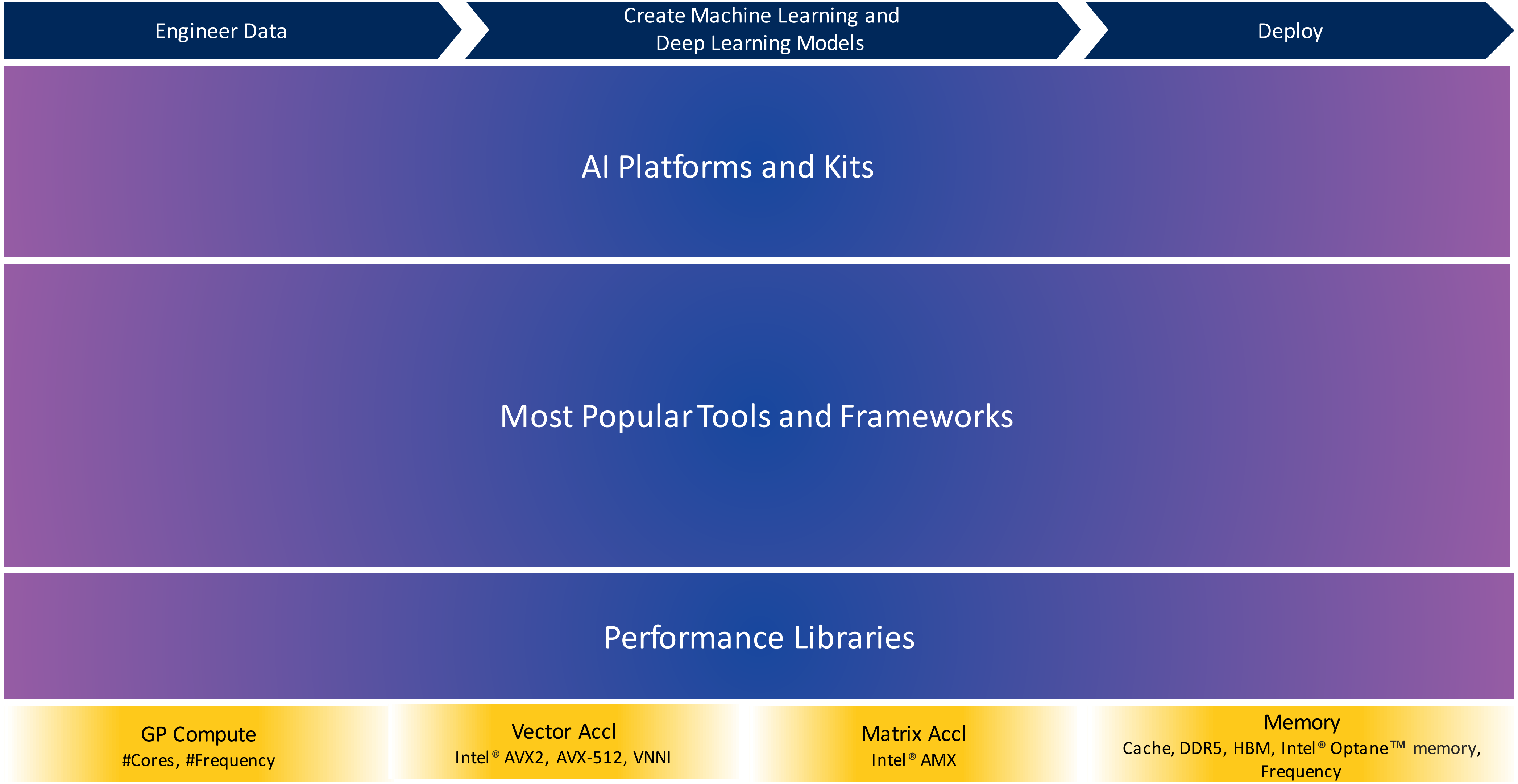
The Habana® Gaudi® AI Training Processor

Gaudi2 outperformed Nvidia A100 MLPerf submissions on both ResNet and BERT
…and First-gen Gaudi achieved near-ideal linear scale on 128- and 256-accelerators


MLPERF ResNet-50 Training Time
(8 accelerator server)


MLPERF BERTTraining Time
(8 accelerator server)

Gaudi2 time-to-train (TTT) improved by 3 to 4.7x compared to first-gen Gaudi



Engineer Data

Create Machine Learning and Deep Learning Models

Deploy

AI Platforms and Kits

Most Popular Tools and Frameworks

Performance Libraries

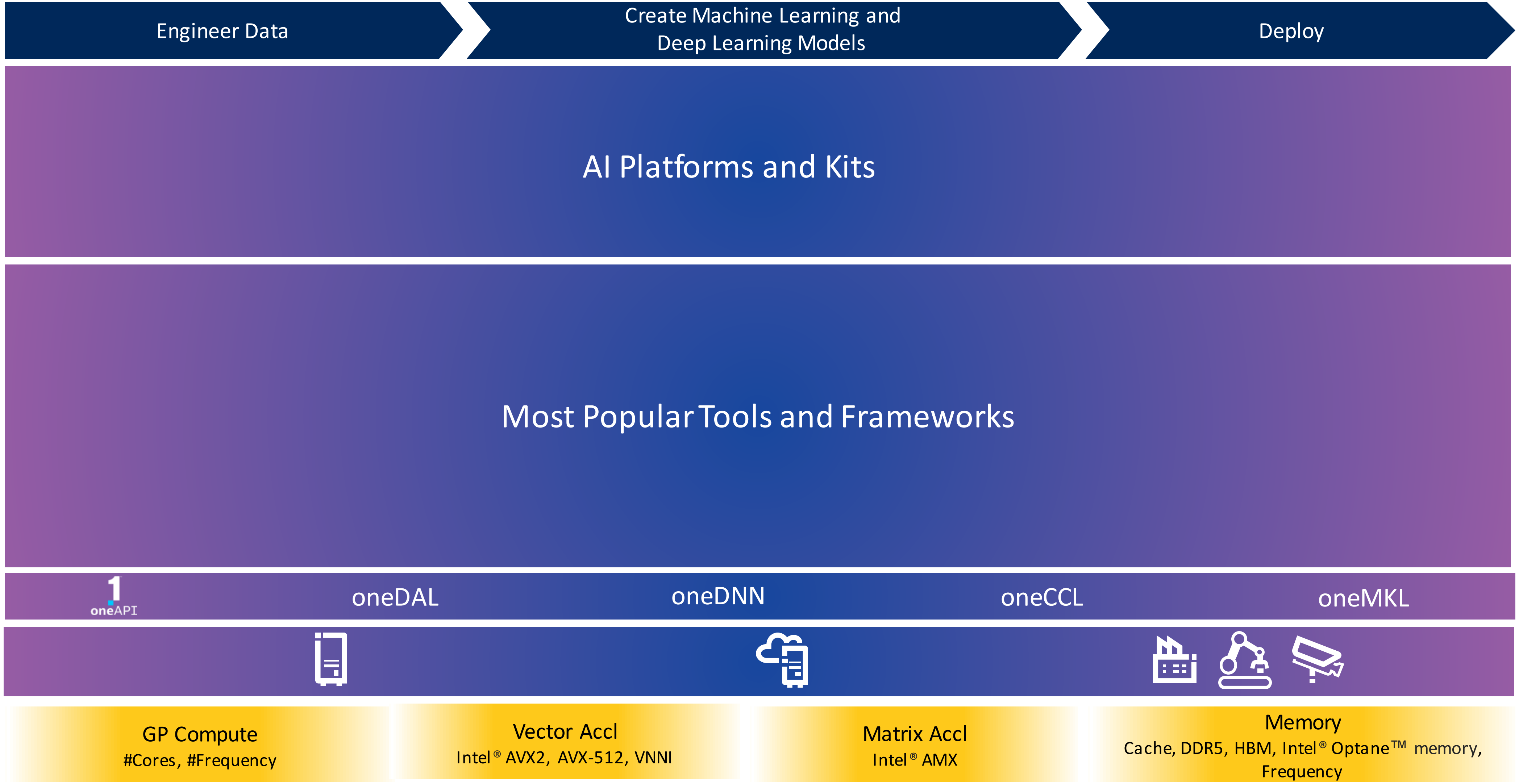
GP Compute
#Cores, #Frequency

Vector Accl
Intel® AVX2, AVX-512, VNNI

Matrix Accl
Intel® AMX

Memory
Cache, DDR5, HBM, Intel® Optane™ memory, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions



Engineer Data

Create Machine Learning and Deep Learning Models

Deploy

AI Platforms and Kits

Most Popular Tools and Frameworks

oneAPI

oneDAL

oneDNN

oneCCL

oneMKL

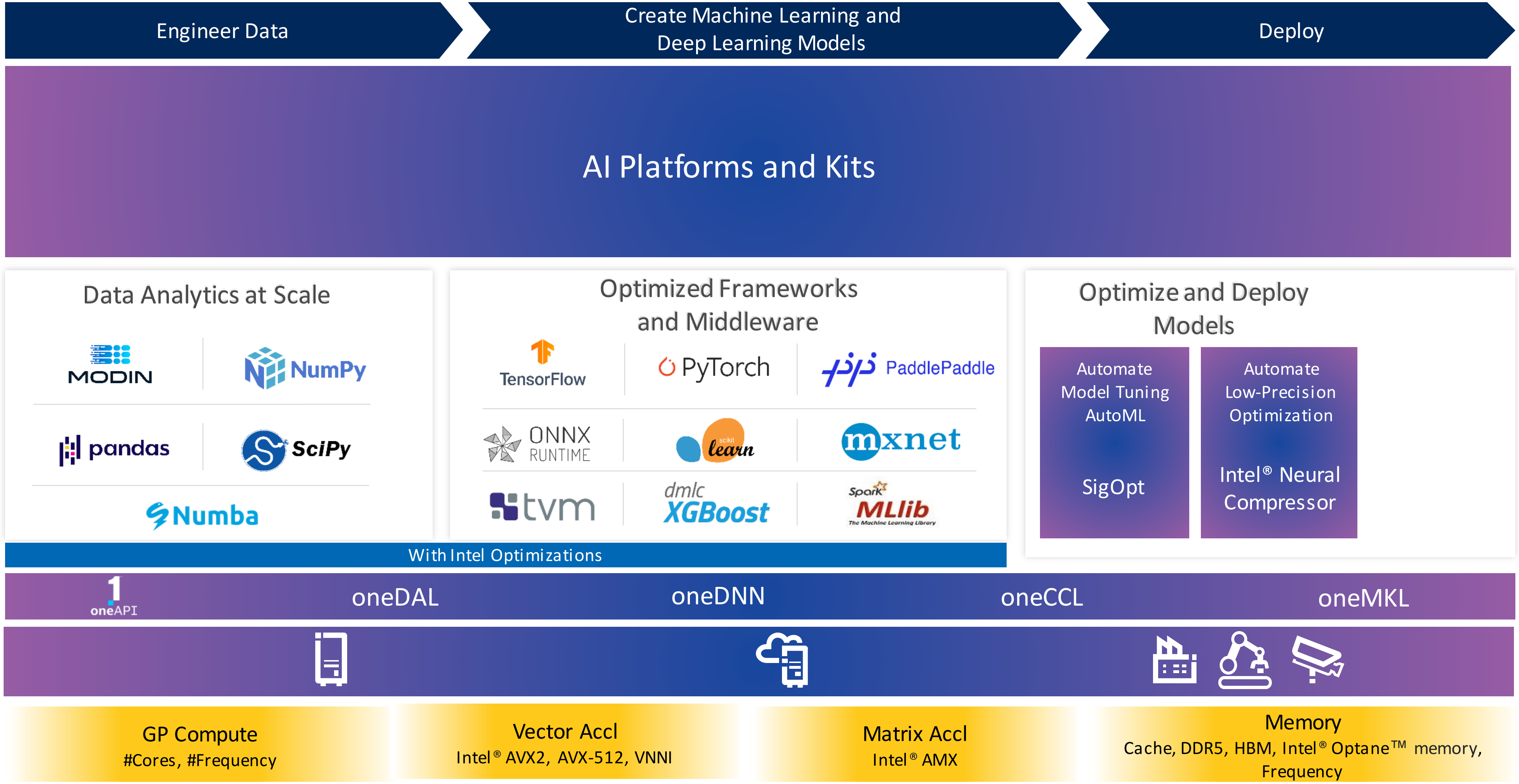
GP Compute
#Cores, #Frequency

Vector Accl
Intel® AVX2, AVX-512, VNNI

Matrix Accl
Intel® AMX

Memory
Cache, DDR5, HBM, Intel® Optane™ memory, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions



Engineer Data

Create Machine Learning and Deep Learning Models

Deploy

AI Platforms and Kits

Data Analytics at Scale

- MODIN
- NumPy
- pandas
- SciPy
- Numba

Optimized Frameworks and Middleware

- TensorFlow
- PyTorch
- PaddlePaddle
- ONNX RUNTIME
- scikit learn
- mxnet
- tvm
- dmlc XGBoost
- Spark MLlib The Machine Learning Library

With Intel Optimizations

Optimize and Deploy Models

- Automate Model Tuning AutoML: SigOpt
- Automate Low-Precision Optimization: Intel® Neural Compressor

oneAPI: oneDAL, oneDNN, oneCCL, oneMKL

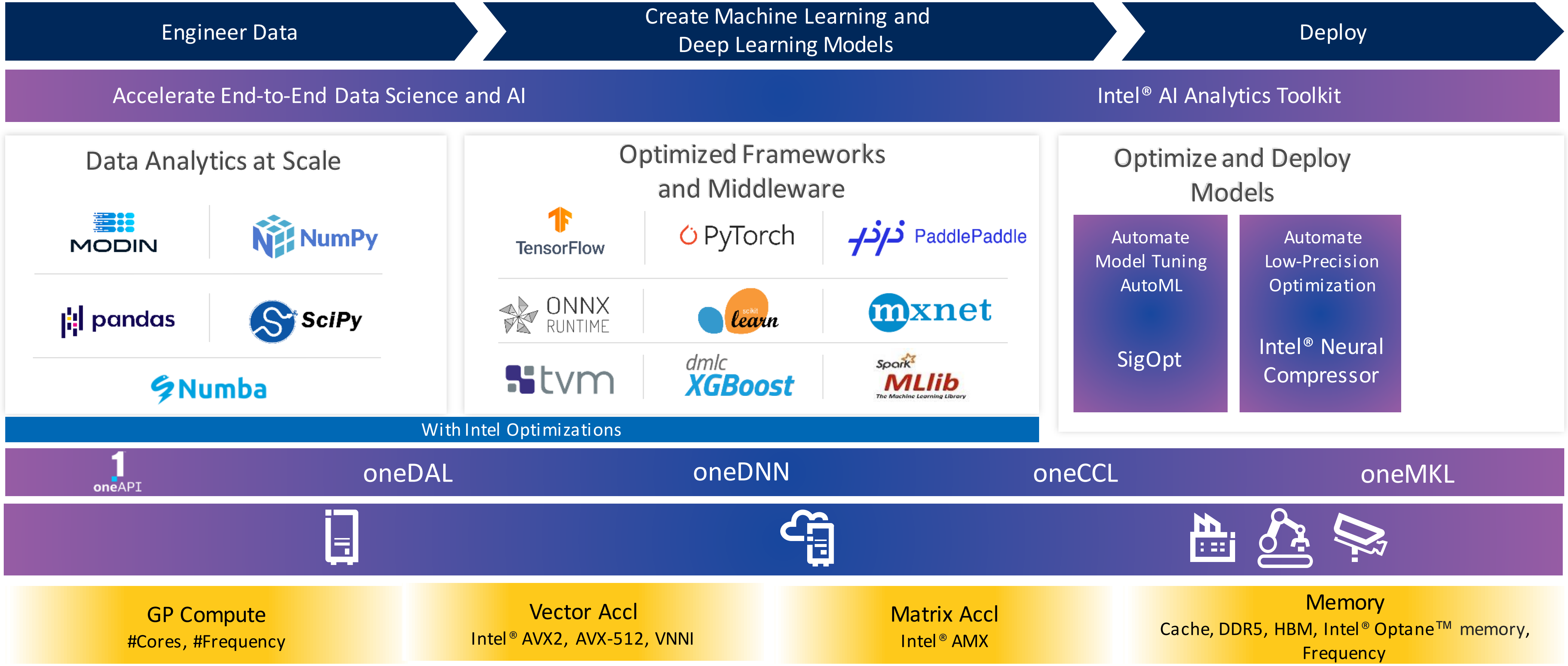
GP Compute
#Cores, #Frequency

Vector Accl
Intel® AVX2, AVX-512, VNNI

Matrix Accl
Intel® AMX

Memory
Cache, DDR5, HBM, Intel® Optane™ memory, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions



Engineer Data

Create Machine Learning and Deep Learning Models

Deploy

Accelerate End-to-End Data Science and AI

Intel® AI Analytics Toolkit

Data Analytics at Scale

MODIN

NumPy

pandas

SciPy

Numba

Optimized Frameworks and Middleware

TensorFlow

PyTorch

PaddlePaddle

ONNX RUNTIME

scikit learn

mxnet

tvm

dmlc XGBoost

Spark MLlib The Machine Learning Library

With Intel Optimizations

Optimize and Deploy Models

Automate Model Tuning AutoML

SigOpt

Automate Low-Precision Optimization

Intel® Neural Compressor

oneAPI

oneDAL

oneDNN

oneCCL

oneMKL

GP Compute
#Cores, #Frequency

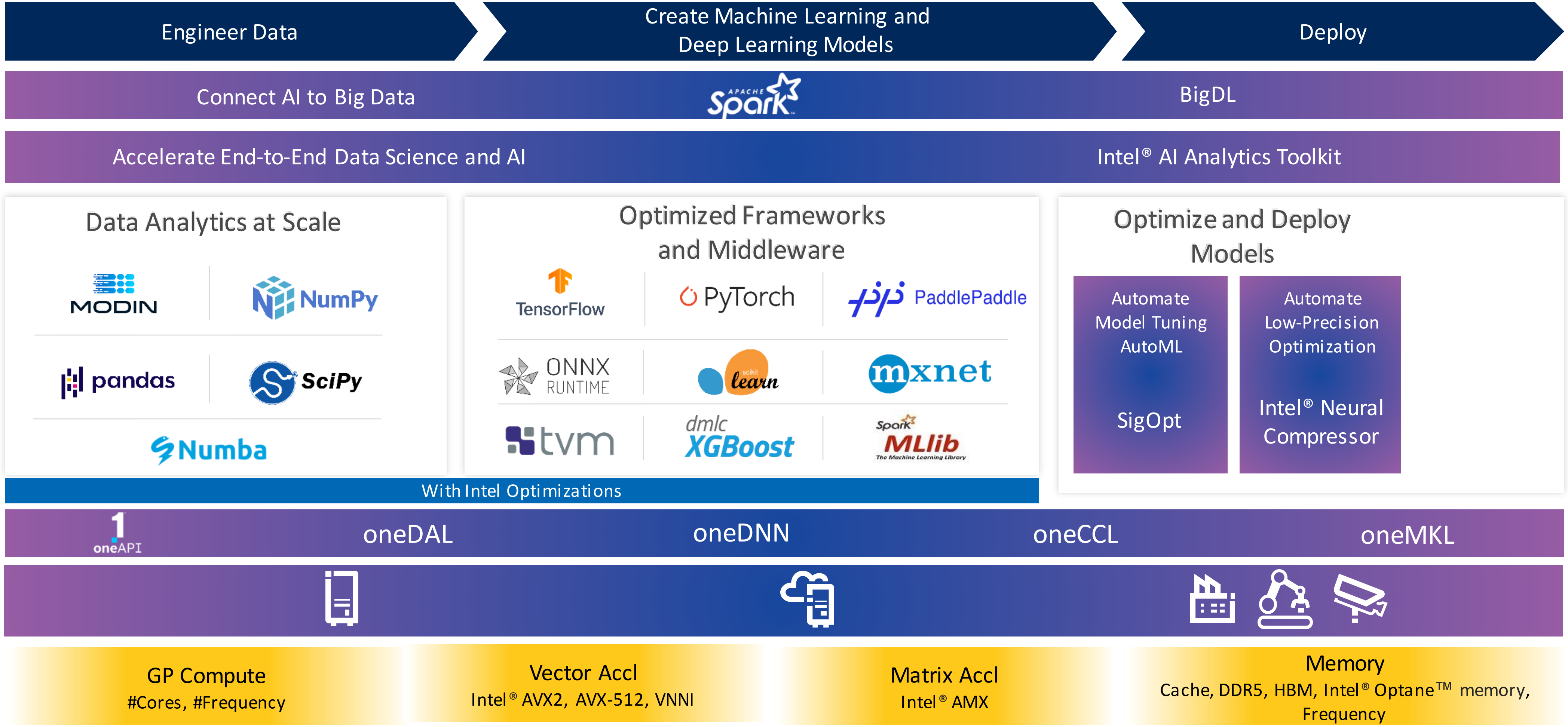
Vector Accl
Intel® AVX2, AVX-512, VNNI

Matrix Accl
Intel® AMX

Memory
Cache, DDR5, HBM, Intel® Optane™ memory, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL-Intel oneAPI Math Kernel Library

AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions



Engineer Data

Create Machine Learning and Deep Learning Models

Deploy

Connect AI to Big Data

Apache Spark

BigDL

Accelerate End-to-End Data Science and AI

Intel® AI Analytics Toolkit

Data Analytics at Scale

MODIN

NumPy

pandas

SciPy

Numba

Optimized Frameworks and Middleware

TensorFlow

PyTorch

PaddlePaddle

ONNX RUNTIME

scikit learn

mxnet

tvm

dmlc XGBoost

Spark MLlib

With Intel Optimizations

Optimize and Deploy Models

Automate Model Tuning AutoML

SigOpt

Automate Low-Precision Optimization

Intel® Neural Compressor

oneAPI

oneDAL

oneDNN

oneCCL

oneMKL

GP Compute

#Cores, #Frequency

Vector Accl

Intel® AVX2, AVX-512, VNNI

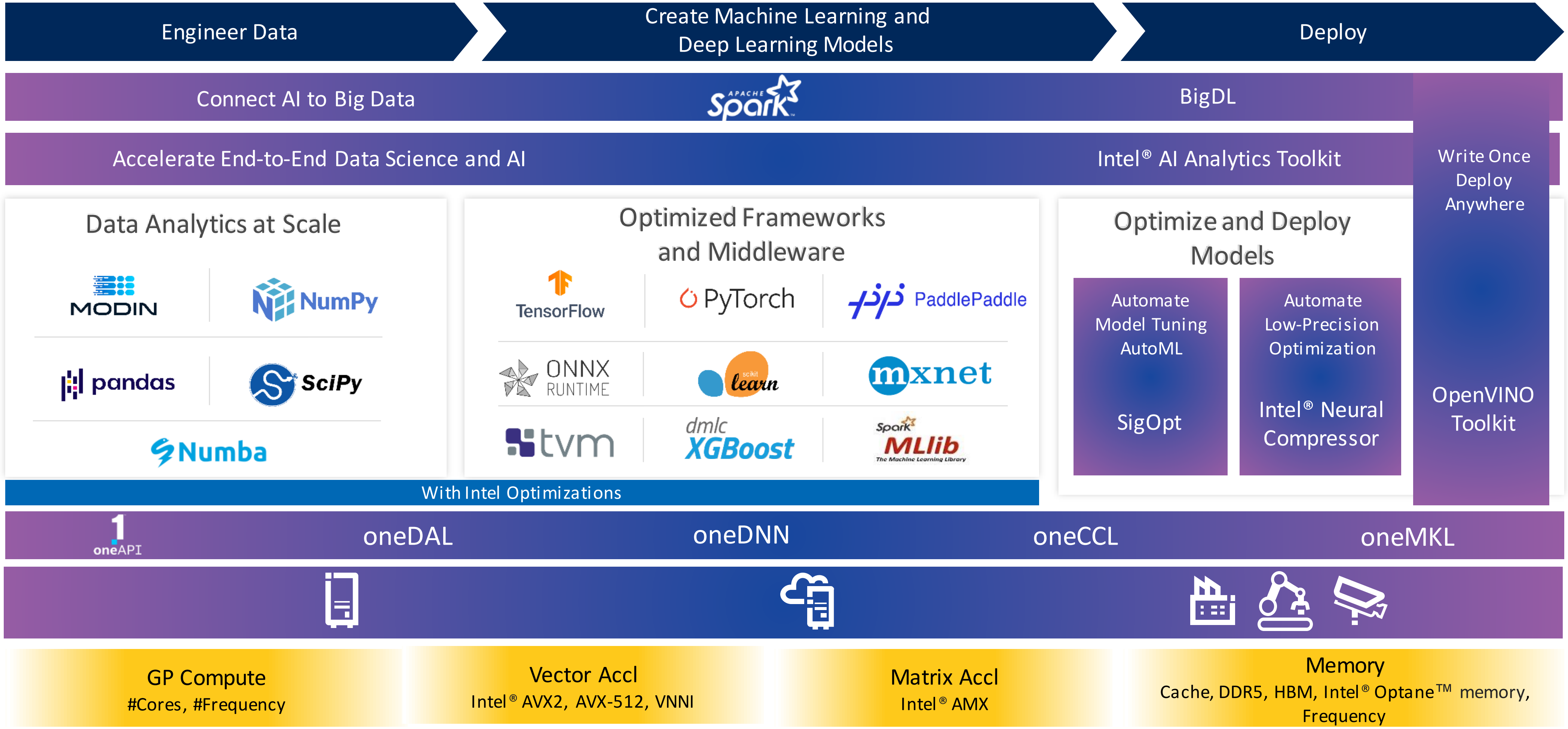
Matrix Accl

Intel® AMX

Memory

Cache, DDR5, HBM, Intel® Optane™ memory, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions



Engineer Data

Create Machine Learning and Deep Learning Models

Deploy

Connect AI to Big Data — Apache Spark — BigDL

Accelerate End-to-End Data Science and AI — Intel® AI Analytics Toolkit

Write Once Deploy Anywhere

Data Analytics at Scale

- MODIN
- NumPy
- pandas
- SciPy
- Numba

Optimized Frameworks and Middleware

- TensorFlow
- PyTorch
- PaddlePaddle
- ONNX RUNTIME
- scikit learn
- mxnet
- tvm
- dmlc XGBoost
- Spark MLlib The Machine Learning Library

With Intel Optimizations

Optimize and Deploy Models

Automate Model Tuning AutoML

SigOpt

Automate Low-Precision Optimization

Intel® Neural Compressor

OpenVINO Toolkit

oneAPI oneDAL oneDNN oneCCL oneMKL

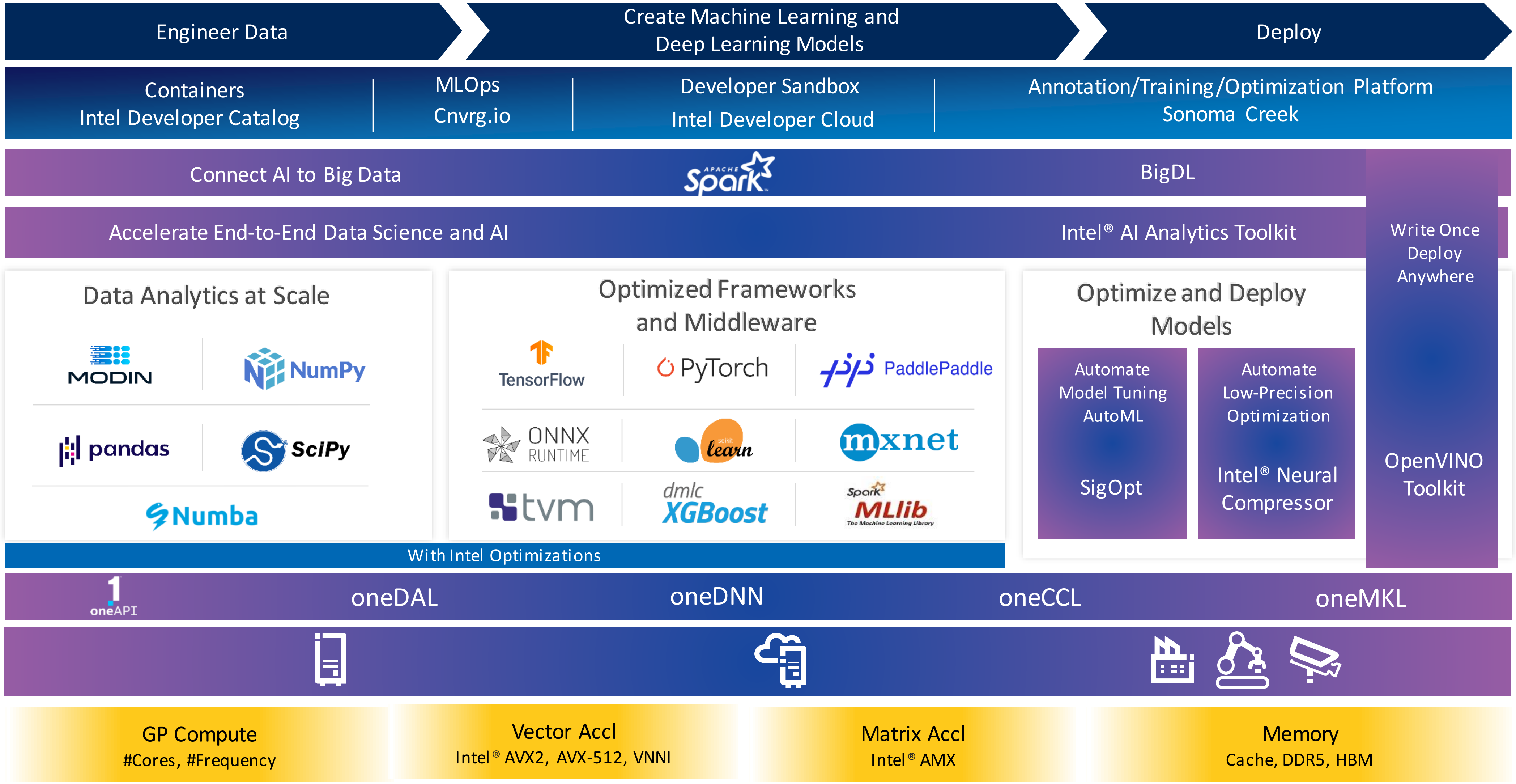
GP Compute
#Cores, #Frequency

Vector Accl
Intel® AVX2, AVX-512, VNNI

Matrix Accl
Intel® AMX

Memory
Cache, DDR5, HBM, Intel® Optane™ memory, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL - Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions



Engineer Data

Create Machine Learning and Deep Learning Models

Deploy

Containers
Intel Developer Catalog

MLOps
Cnvrg.io

Developer Sandbox
Intel Developer Cloud

Annotation/Training/Optimization Platform
Sonoma Creek

Connect AI to Big Data

Apache Spark

BigDL

Accelerate End-to-End Data Science and AI

Intel® AI Analytics Toolkit

Write Once
Deploy
Anywhere

OpenVINO
Toolkit

Data Analytics at Scale

MODIN

NumPy

pandas

SciPy

Numba

Optimized Frameworks and Middleware

TensorFlow

PyTorch

PaddlePaddle

ONNX RUNTIME

scikit learn

mxnet

tvm

dmlc XGBoost

Spark MLlib

Optimize and Deploy Models

Automate
Model Tuning
AutoML

SigOpt

Automate
Low-Precision
Optimization

Intel® Neural Compressor

With Intel Optimizations

oneAPI

oneDAL

oneDNN

oneCCL

oneMKL

GP Compute
#Cores, #Frequency

Vector Accl
Intel® AVX2, AVX-512, VNNI

Matrix Accl
Intel® AMX

Memory
Cache, DDR5, HBM

oneDAL–InteloneAPIData AnalyticsLibrary,oneDNN–InteloneAPIDeepNeuralNetworksLibrary,oneCCL–InteloneAPICollectiveCommunications Library,oneMKL-InteloneAPIMathKernelLibrary

AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

Red Hat OpenShift Data Science

Tools and capabilities

Jupyter notebooks

Conduct exploratory data science in JupyterLab with access to core AI/ML libraries and frameworks including TensorFlow and PyTorch using our notebook images or your own.

Source-to-image (S2I)

Publish models as end points via S2I for integration into intelligent apps. Rebuild and redeploy based on changes to the source code.

GPU Acceleration

Accelerate your data science experiments through the use of GPU acceleration on the Red Hat OpenShift Dedicated platform.

Building on the foundations of data science

Red Hat OpenShift

Key Features of Red Hat OpenShift Data Science

Addressing AI/ML experimentation and integration use cases on a managed platform

Cloud Service

Available on Red Hat OpenShift Dedicated (AWS) and Red Hat OpenShift Service on AWS

Increased capabilities/collaboration

Combines Red Hat components, open source software, and ISV certified software available on Red Hat Marketplace

Core data science workflow

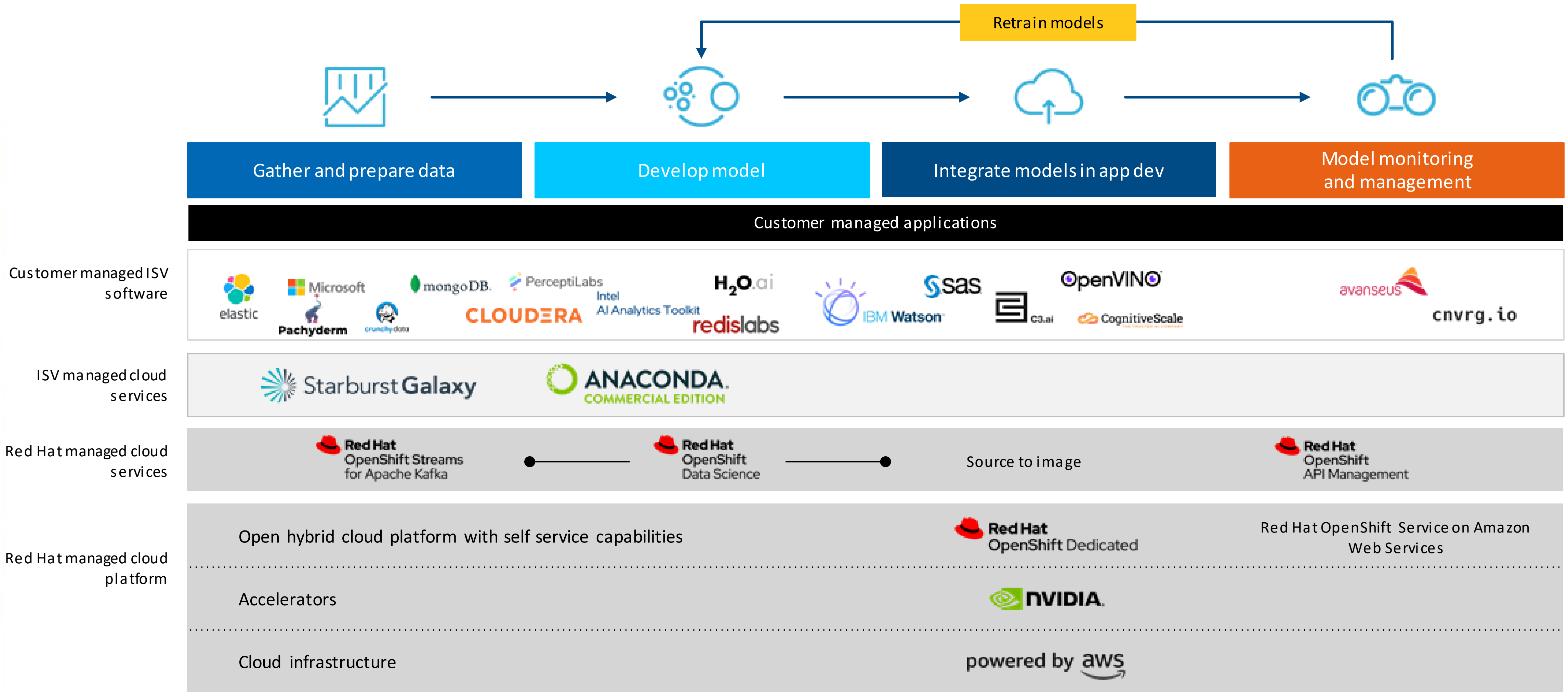
Provides data scientists and intelligent application developers the ability to build, train, and deploy ML models

Rapid experimentation use cases

Model outputs are hosted on the Red Hat OpenShift managed service or exported for integration into an intelligent application

Red Hat
OpenShift

… and Integrating our Partner Ecosystem



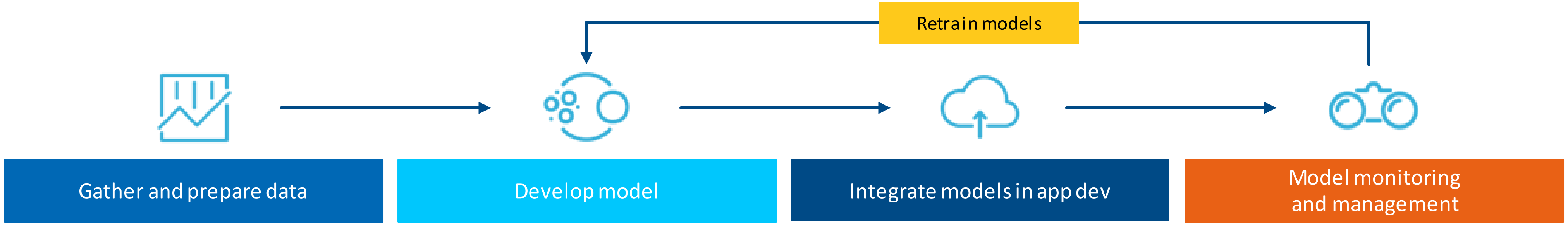
Retrain models

Gather and prepare data → Develop model → Integrate models in app dev → Model monitoring and management

Customer managed applications

Customer managed ISV software	elastic, Microsoft, Pachyderm, crunchy data, mongoDB, PerceptiLabs, CLOUDERA, Intel AI Analytics Toolkit, H2O.ai, redislabs, IBM Watson, sas, C3.ai, OpenVINO, CognitiveScale, avanseus, cnvrg.io
ISV managed cloud services	Starburst Galaxy, ANACONDA COMMERCIAL EDITION
Red Hat managed cloud services	Red Hat OpenShift Streams for Apache Kafka — Red Hat OpenShift Data Science — Source to image; Red Hat OpenShift API Management
Red Hat managed cloud platform	Open hybrid cloud platform with self service capabilities: Red Hat OpenShift Dedicated; Red Hat OpenShift Service on Amazon Web Services Accelerators: NVIDIA Cloud infrastructure: powered by aws

Red Hat OpenShift Data Science + Intel® AI



Customer managed applications

Intel® Software Components

Coming soon cnvrg.io

DATASETS EXPERIMENT BUILD DEPLOY MONITOR

intel. AI ANALYTICS TOOLKIT

Synapse AI Coming soon

intel OpenVINO

Red Hat managed cloud services

Red Hat OpenShift Streams for Apache Kafka

Red Hat OpenShift Data Science
TensorFlow jupyter PyTorch

Source to image

Red Hat OpenShift API Management

Red Hat managed cloud platform

Open hybrid cloud platform with self service capabilities — Red Hat OpenShift Dedicated — Red Hat OpenShift Service on Amazon Web Services

Intel® Accelerators — intel. XEON — Coming soon habana DL1

Cloud infrastructure — powered by aws

Demo: Red Hat OpenShift Data Science + Intel® AI

On-Prem AI Solution

A turn-key AI system solution that allows the data scientist to only focus on their model building and training while allowing IT to forget about the underlying complex infrastructure, scaling challenges, and cost of iterating.

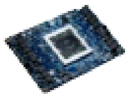
cnvrg.io

Datasets | Experiment | Build | Deploy | Monitor

intel. oneAPI AI ANALYTICS TOOLKIT | OpenVINO™ | SynapseAI® | TensorFlow

Red Hat OpenShift | Red Hat Enterprise Linux

SUPERMICRO



Available in Q4!

Expansion SKUs can be ICX, Gaudi® or a DDN Box

Expansion that Scales to Any Size Needed

SUPERMICRO

Expansion SKUs

Base SKU

Common Issues in Machine Learning

Our AI projects are fragmented, take forever, and don't deliver what we expected

The data science team has their own process and stacks outside of our other development stacks and workflows

Coordinating data scientists, dev, security, and ops takes too long

It takes too long to stand up compute servers with specialized hardware

Can Kubernetes help us move AI workloads across data centers and clouds?

We're using files and folders to manage datasets, code, models, and metadata

Can we manage AI at scale the way we manage other software projects?

cnvrg.io: Operating System for AI

Everything needed to build and deploy AI on any infrastructure

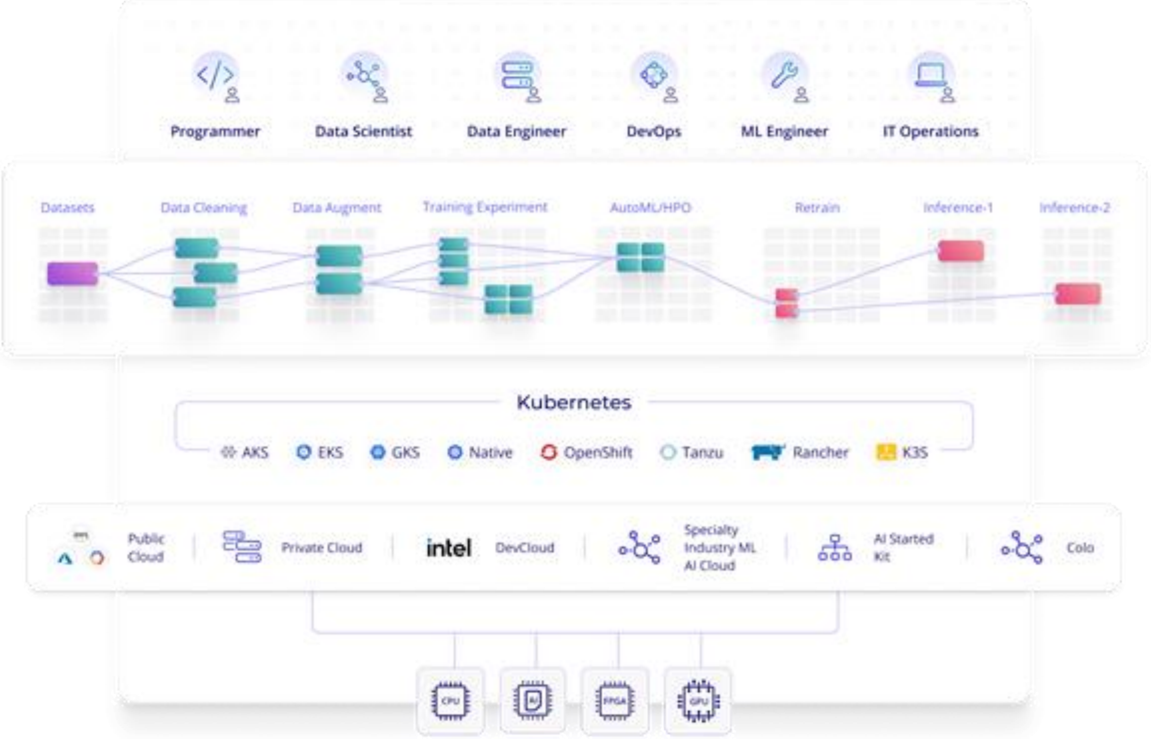
Control Plane
Management layer for datasets, model code, jobs, model performance, cluster, and resource statistics

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

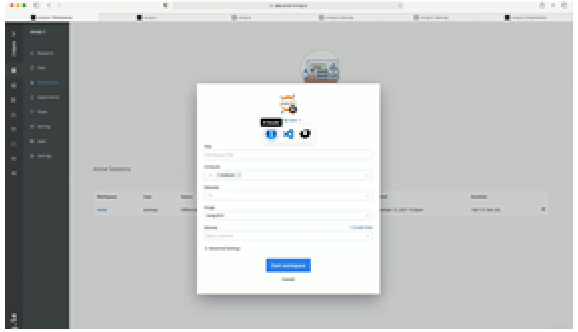
Orchestration and Scheduling
Kubernetes-based meta-scheduler for orchestration, scheduling, and scaling across clusters

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

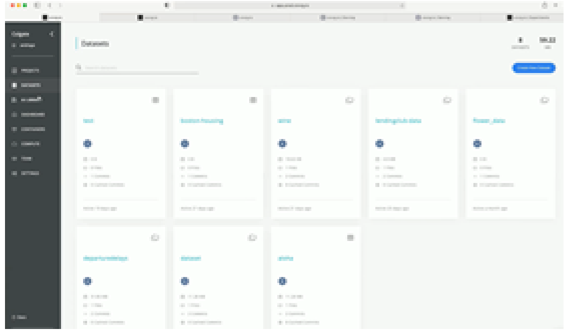


cnvrg Simplifies ML Workflows from End-to-End

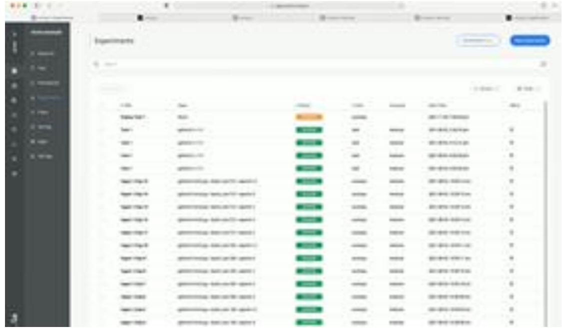
1 Create projects and workspaces



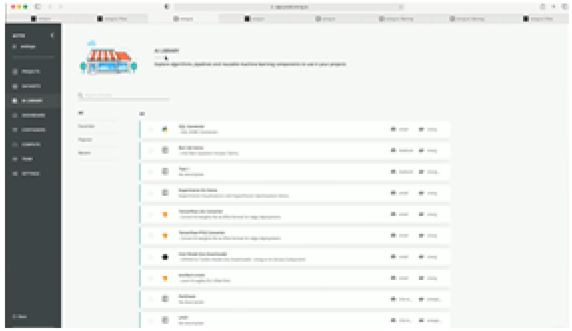
2 Connect data



3 Manage experiments

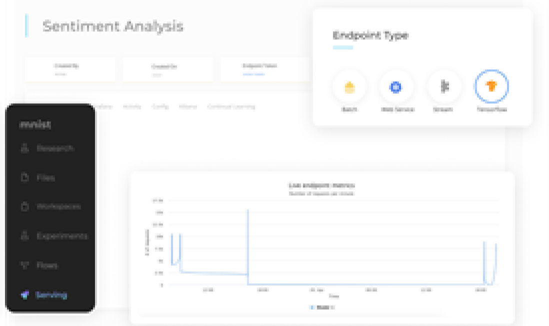


4 Create and re-use models



5 Drag-and-drop ML pipelines

6 Deploy and monitor models/clusters



Demo

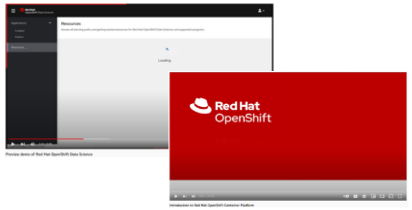
Red Hat OpenShift - cnvrg –AI Toolkit – openVINO - Habana

Checkout Intel and Red Hat AI Developer Program

Go to the Intel and Red Hat AI Developer Program:
https://www.intel.com/content/www/us/en/developer/partner/overview.html

How-To Videos

Videos showing developer experience with OpenShift, RHODS, cnvrg.io, AI Toolkit & OpenVINO

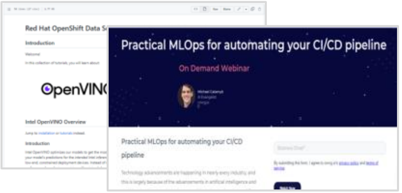


Sandbox Integration

RHODS Sandbox, cnvrg.io Metacloud

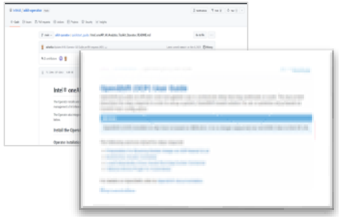
Learning Pathways

Pathways for AI Toolkit and OpenVINO, Webinars and Workshops for cnvrg.io


Red Hat OpenShift Data S
OpenVINO
Practical MLOps for automating your CI/CD pipeline
On Demand Webinar
Practical MLOps for automating your CI/CD pipeline

Quick Start Guides

How to get up and running



Red Hat OpenShift Data Science

AI ANALYTICS TOOLKIT

OpenVINO

cnvrg.io

habana An Intel Company

Notices and Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:

From the landing page, go to the Events tab and then to Intel Innovation 2022.

intel®
innovation

Thank you