



AI Hype: Learn how your business can surf this wave successfully

Jan Krüger
Intel Corp

SAPinsider
2023



In This Session

You will hear from Intel experts why Large Language Models like Microsoft OpenAI have become such industry hype, and how you can benefit from the constant innovation with Intel Solutions, ranging from hardware products to software and solutions.

Agenda

- AI Hype vs. AI Help?
- AI Software Ecosystem is more as NVidia
- Intel AI Software & Intel Development Cloud
- Optimize Performance to reduce Costs
- Wrap-Up



AI Hype vs. AI Help?

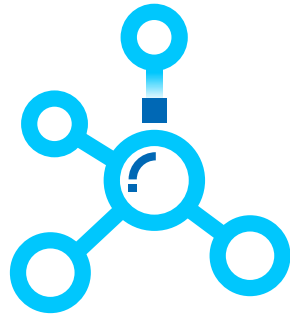
- AI means different things to different people.
- Lets focus:
 - On Traditional Maschine Learning (ML)
 - Deep Learning (DL)
- And within Deep Learning:
 - on „Large“ Language Models (LLM) like ChatGTP
 - on AI pipelines for DL Training
 - and the different needs for DL Inference & DL Training

LLMs in Enterprises?

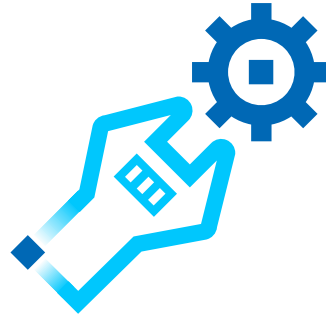
- LLMs are unregulated as of today
- Using SAP, Microsoft, Google, X or others LLMs potentially leads to IP leakage and using the output is not IP protected either.
- However, LLMs hosted and trained in „Private“ environment unleash various opportunities
- Intel embraces AI Everywhere, and enables Enterprise to deploy „Private“ environments or protect data used with 3rd party LLMs.

Intel's AI Software Objective – Consumable and Easy to Use

Simplify our AI developers' lives regardless of how they consume the software



Open source, common software programming model

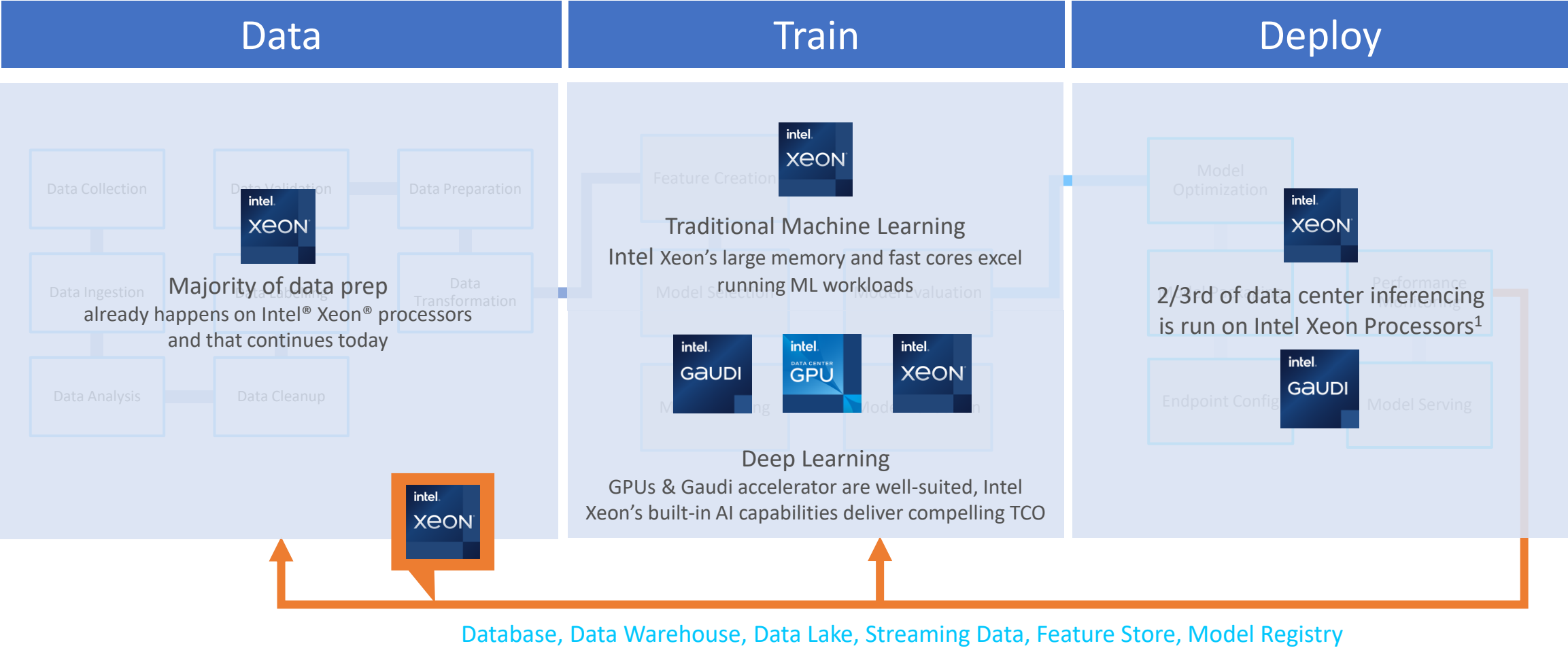


End-to-end tools and kits to help accelerate time-to-solution



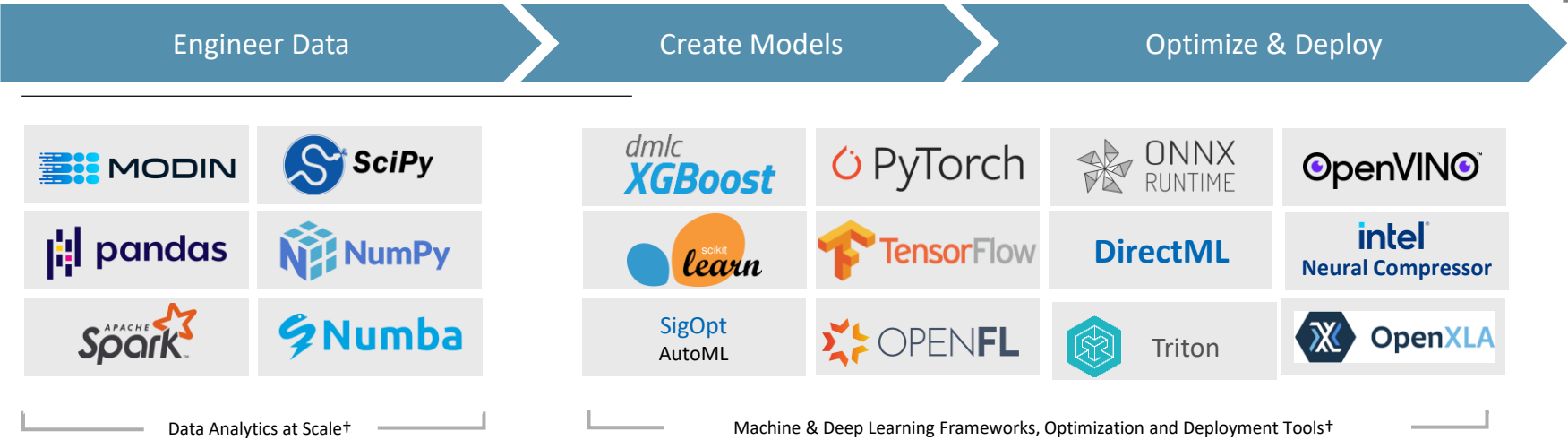
Meet developers where and how they use software

The AI Pipeline Runs on Intel




¹ Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2022.

Intel AI Software Ecosystem Approach, Open & Scalable




Intel® Developer Cloud and
Intel® Developer Catalog
Try the latest Intel tools and hardware, and
access optimized AI Models


cnvrg.io
Full stack ML operating system


Hugging Face
Intel optimizations and fine-tuning recipes, optimized
inference models, and model serving

Note: components at each layer of the stack are optimized for targeted components at other layers based on expected AI usage models, and not every component is utilized by the solutions in the rightmost column

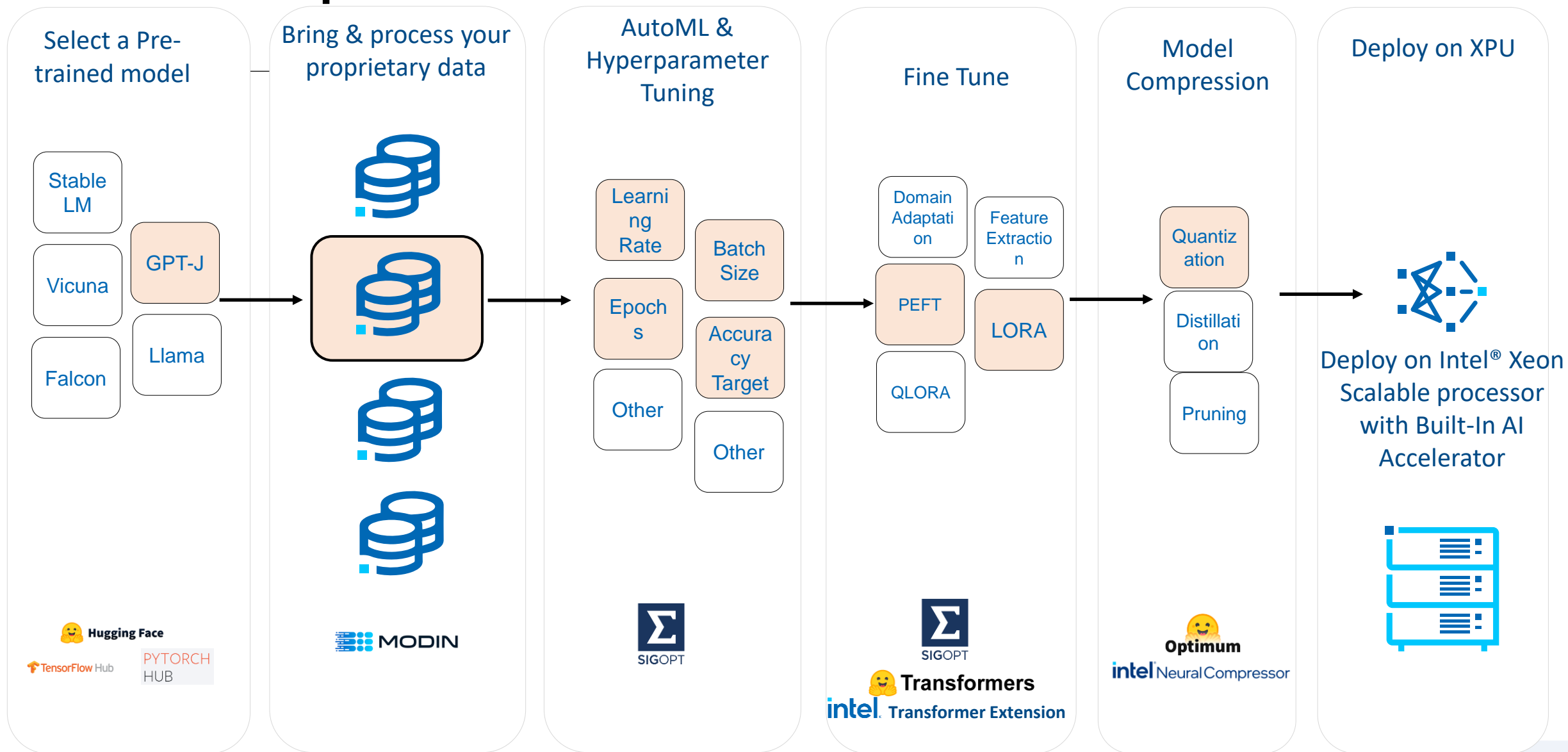
† This list includes popular open source frameworks that are optimized for Intel hardware

Intel® Xeon® Scalable Processor Software Value Approach Map

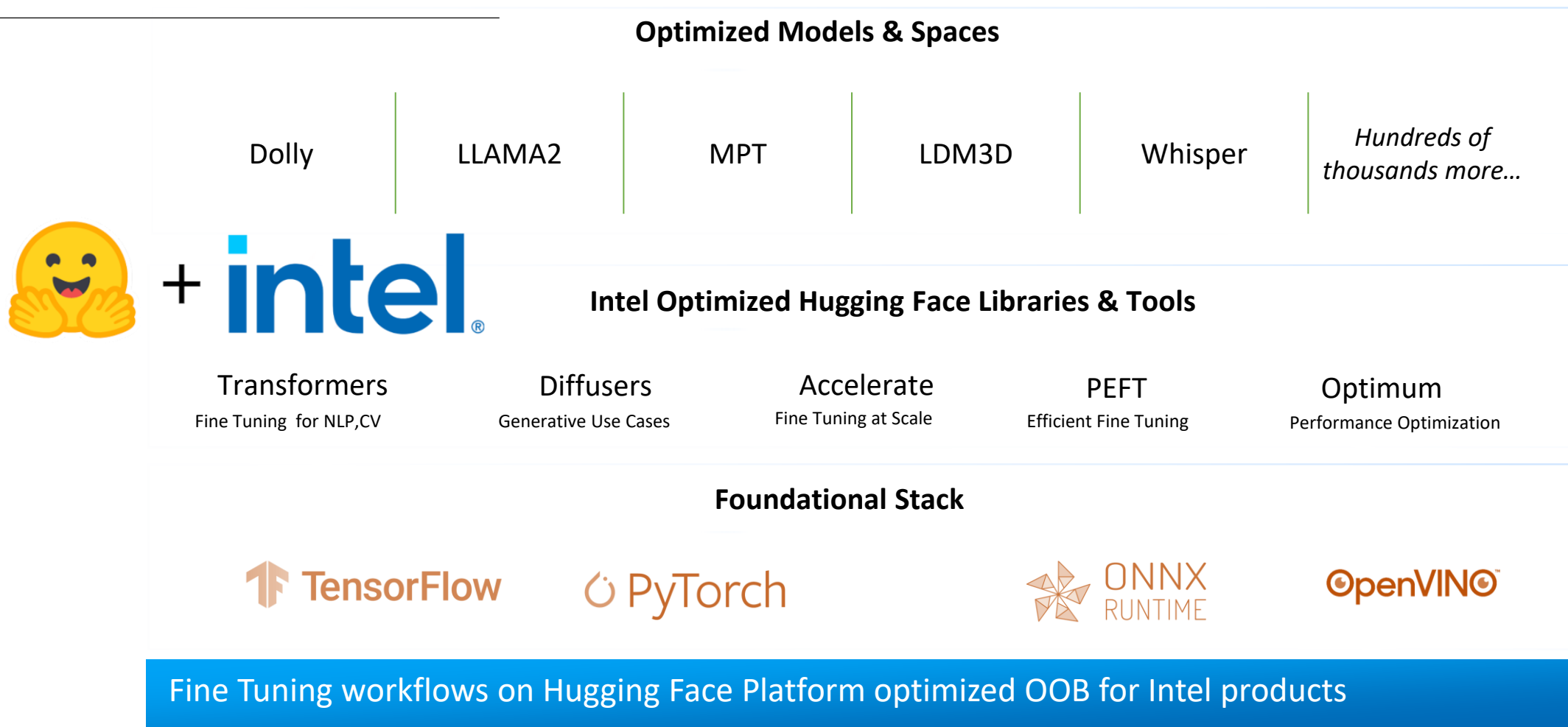
Categories	SW Product	Optimizations Upstreamed	Intel Extension	Intel Distro
DL Frameworks	TensorFlow	■ oneDNN integrated. Up to 3x performance boost	■ Add'l perf enhancements	
	PyTorch	Differentiation upstreamed to Open Source	+ Differentiation in Intel package	
ML Frameworks	Scikit-Learn			
Data Prep	Spark / OAP			
	Modin			

See link below for workloads and configurations. Results may vary
<https://www.intel.com/content/www/us/en/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html>

LLMs in Enterprise with Intel



In Focus: HuggingFace and Intel



<https://huggingface.co/Intel>

Intel Developer Cloud

- Start your AI, ML, LLM usage today
- Easy to use, fast time to market
- Secure and Resilient



Intel® Developer Cloud

Developers

Provide an **easy path for developers** to access and use Intel-optimized AI software supported by Intel's latest Xeon and GPU compute.

Companies

Offer a **delivery vehicle** to bring new Intel technologies and products to market (faster) and a platform to **monetize AI services**.

Partners

Provide performance and **cost-optimized AI compute services** to internal and external AI SaaS providers

- **Ready access to Intel AI Portfolio**
 - Evaluate, prototype, validate perf on 4th Gen Intel® Xeon processors with AMX, Intel® Gaudi2® accelerators, Intel® Max Series CPUs & GPUs and Intel® Flex Series GPUs
- **Increased productivity**
 - Build on top of GAI/LLM SW stack pre-optimized for Intel on all major ML frameworks and popular data science libraries
 - Start from Intel optimized Model zoo of popular models, fine tune to domain/customer specific datasets
 - No lock-in, same models can be used on-prem or in 3rd party CSP
- **Accelerated TTM & Scale**
 - Agility to adapt to rapidly evolving technology landscape with new models/services – LLMs, GAI, multi-modal models
 - Choose the right sized platform according to AI phase requirement, training, fine tuning, transfer learning or inferencing

Intel® Developer Cloud

AI Infrastructure Services

Service	AI Services		
Use case	Deploy AI workloads on Intel platforms		
Access	SSH / CLI k8s Cluster API		
Software	Ubuntu OS Latest Intel kernel drivers Intel (optimized) AI frameworks		
Runtime Environment(s)	<div><div>Dedicated host Linux OS</div><div>VM</div><div>k8s</div></div>		
Hardware	<div><div>intel. XEON</div><div>intel. GAUDI</div><div>intel. DATA CENTER GPU MAX SERIES</div><div>intel. DATA CENTER GPU FLEX SERIES</div></div>		

Typical use case(s):

- LLM model training and optimization
- AI model deployment for inferencing

Usage model

- AI model deployment via CLI/SSH automation
- AI container deployment via k8s APIs

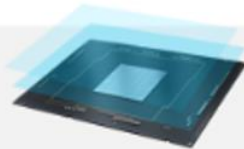


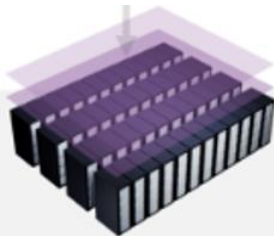
Target customers

- AI disruptors (startups) & partners
- Established AI-savvy enterprises

Examples

- Training or Finetuning of large AI model
- Deploying inferencing workload
- Deploying AlaaS offerings based on Intel AI portfolio

Intel® Developer Cloud AI Compute Instances

Use Case	Small model Inferencing	Small model Inferencing & Fine-tuning	Large model Inferencing Small model Training & Fine-tuning	Large model Training
Instance Type	 Medium VM Xeon	 Large VM Xeon or PVC card	 Full System Gaudi2 or PVC system	 Full Cluster Multiple Gaudi2 systems
Scale	Up to 7 billion parameters	Up to 20 billion parameters	Up to 50 billion parameters	Hundreds of billions of parameters
Milestones	<ul style="list-style-type: none">• Xeon-based VMs with AMX – Nov '23	<ul style="list-style-type: none">• Xeon PVC-enabled VMs – Nov '23	<ul style="list-style-type: none">✓ Dedicated multi-card Gaudi2 & PVC systems - available today	<ul style="list-style-type: none">• Gaudi2 clusters – available in Q4 '23

Optimize TCO



Wrap up

Where to Find More Information

- Intel Developer Cloud <https://devcloud.intel.com>

Key Points to Take Home

Jan.Krueger@intel.com

Please remember to complete
your session evaluation.

SAPinsider



SAPinsider.org

PO Box 982Hampstead, NH 03841
Copyright © 2023 Wellesley Information Services.
All rights reserved.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies. Wellesley Information Services is neither owned nor controlled by SAP SE.

SAPinsider comprises the largest and fastest growing SAP membership group worldwide, with more than 750,000 global members.
