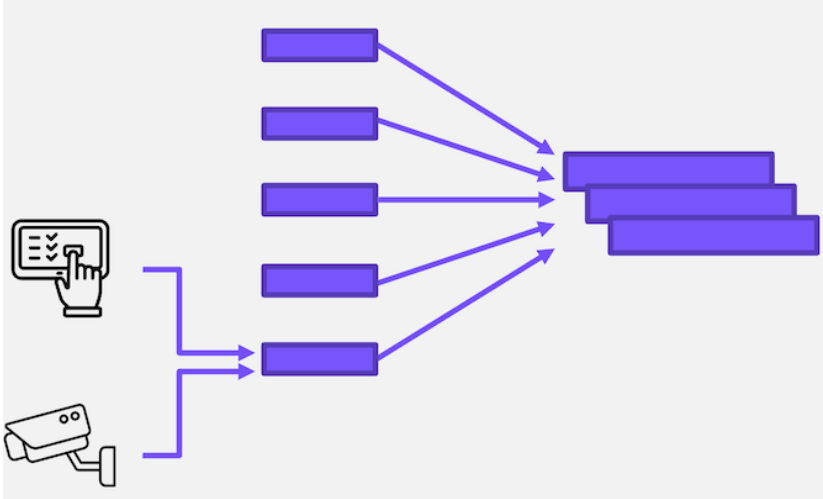
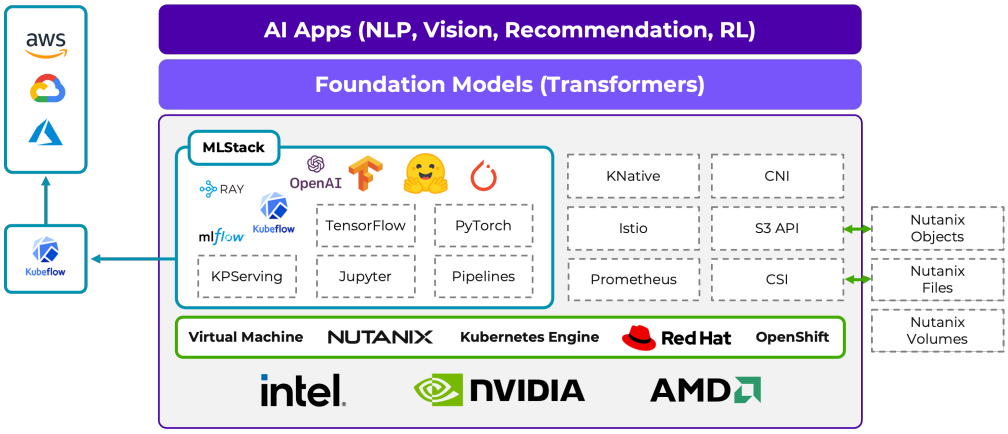
Book of AI/ML - AI on Nutanix

[PDF generated August 17 2023. For all recent updates please see the Nutanix Bible releases notes located at https://nutanixbible.com/release_notes.html. Disclaimer: Downloaded PDFs may not always contain the latest information.]

There is a growing need for infrastructure that can span across edge, core, and/or the public cloud. For example, data can be generated and processed locally at several far-edge nodes, then sent back to the core datacenter for aggregate decision-making and fine-tuning the model.



Regardless of what is deployed, you need a consistent operating model across the tiers. Nutanix Cloud Platform can run at all locations - whether it's a single-node edge device, a multi-node cluster in your core datacenter, or in the public cloud with NC2. This provides a unified cloud operating model that enables the simple, consistent operation of your AI/ML platform.



NCP supports a variety of hardware and NVIDIA GPU cards, and you can run your workloads on Nutanix Kubernetes Engine, Red Hat Openshift, or on VMs.

A key aspect of running AI/ML workloads is managing the machine learning lifecycle. The term MLOps is a compound of Machine Learning and Operations, and is analogous to DevOps, in that it aims to automate, orchestrate, and deploy models in a consistent and efficient manner. Kubeflow is a popular solution for MLOps, and the Nutanix Cloud Platform is validated for Kubeflow, enabling consistent MLOps across the tiers.

For more information on running AI/ML on Nutanix, check out the following resources:

- Nutanix.com AI/ML Solutions Page
- Nutanix Validated Design for the AI-enabled Edge
- Nutanix.dev blog posts
- DEMO: Running a Chatbot on Nutanix Cloud Platform