Book of AI/ML

AI is used almost everywhere in our daily lives. For example, your smartphone can recognize images and voices, translate languages, identify music, and more. What about AI in your datacenter? How can it help you, and more importantly, how can you get there?

The enterprise data explosion - all of the databases, conversations, logs, metrics, etc. that get generated in an organization - is fueling the demand for AI to extract actionable insights for making better business decisions. All industries can benefit from actionable insight generation from AI. Until 2021, most enterprise AI applications were discriminative in nature, focusing on the computation of a decision boundary for classification (for example, text or speech recognition, object detection) and regression tasks (for example, sales forecasting, stock price prediction). Since 2022, there has been a sweeping change in the enterprise AI landscape with increasing adoption of generative AI-based applications triggered by the popularity of chatGPT and diffusion models. Unlike discriminative AI, generative AI generates a distribution from sample data (i.e. it creates new data).

The Nutanix Cloud Platform provides the ideal stack for running your AI/ML workloads, especially for generative AI applications, which take massive computational resources. The following chapters will cover this in more detail.