Big Book of MLOps

Jamie Ralph

2024-09-24

Table of contents

Preface

Machine learning operations (MLOps) is a set of practices that applies DevOps principles to the ML model lifecycle. MLOps teams manage a complex ecosystem of tools and platforms that data scientists and ML engineers use to productionise their models. I started working in MLOps in February 2024 and quickly found it a steep learning curve. One of the challenges was just knowing where to start and where to look for information. This book tries to bring together the resources I've found most useful for working in MLOps.

This book is a work in progress. The planned sections for the book fall roughly into these categories:

- Kubernetes
- Linux
- Python
- Essental non-technical skills (e.g. time management)

Networking in Kubernetes

Networking is a fundamental component of a Kubernetes cluster. There are many networking requirements within a cluster: containers communicate with each other in a pod, pods communicate with each other (both on the same node and across different nodes), and traffic from outside the cluster must reach the correct pod. Pods are ephemeral (they are destroyed and recreated) and have changing IP addresses, so how do we handle this? The resources here give a good overview of how Kubernetes networking is implemented.

Resources on Kubernetes networking

- Services, Load Balancing, Networking from the official Kubernetes docs gives a nice overview of Kubernetes networking APIs and requirements.

- The official docs on Services goes into more detail on how services enable reliable communication with groups of pods whose IP addresses change as they're destroyed and recreated.
- This article on different service types by KodeKloud goes into detail about the different service types available in Kubernetes.
- Tracing the path of network traffic in Kubernetes by Kristijan Mitevski goes into much more detail about the mechanisms underlying pod to pod communication. The author covers key concepts like network namespaces, interfaces, network switches, and container network interfaces (CNI).
- The official docs on EndpointSlices shows how Kubernetes bridges the gap between services and pods.
- This video on Kubernetes ingress by TechWorld with Nana gives a really nice overview of how Kubernetes services are exposed securely to the outside world. The video includes details on routing, ingress controllers, enabling HTTPS, and how ingress looks different in cloud platforms vs bare metal.
- Kubernetes uses DNS to allow workloads to find services using consistent domain names instead of IP address. Tech Tutorials with Piyush introduces this topic with a hands-on demo. The official docs also talk about how DNS works in Kubernetes and provides a comprehensive guide to debugging DNS.

Resources for networking on bare metal cluster configurations

- MetalLB is a load-balancer implementation for bare metal Kubernetes clusters. It enables administrators of bare metal clusters to expose services using type LoadBalancer. This video by Engineering with Morris provides a clear and concise explanation of how MetalLB works.

Resources for service meshes

- A service mesh is a piece of infrastructure that makes communication between Kubernetes microservices more secure and traceable. There are different service meshes available, one of which is **istio**. This video from Tech with Nana discusses how istio works and what problems it solves.

Resources on proxies and reverse proxies

- What is a reverse proxy? by CloudFlare provides a good overview of why we need proxies and reverse proxies for network communication
- NGINX is a popular reverse proxy - this introduction to NGINX by Sanjeev Sharma provides a hands-on demo of setting up NGINX and testing its capabilities

Infrastructure as code

Infrastructure as code (Iac) is a way to provision and manage infrastructure using code, rather than manual processes. IaC tools help to automate infrastructure management, and allow teams to track changes in infrastructure using version control.

Resources on Ansible

- Ansible is a tool for automating IT infrastructure. It combines pre-built modules and declarative **playbooks** to automate common administrative tasks on Linux servers. This video series from Learn Linux TV is a great introduction to Ansible.
- Kubespray is a tool that uses Ansible to manage Kubernetes clusters (e.g. deploying a cluster, adding/removing nodes). This video from Engineering with Morris runs through the setup process for a homelab Kubernetes cluster using Kubespray.

Architecture patterns with Python

Software architecture is an important concept for writing software systems that are robust and easily maintained. While working in MLOps doesn't often require writing entire software systems, I've found that architecture patterns have helped me think about ways to design better automation scripts.

Resources on architecture patterns in Python

- By far the best resource I've found on this topic is Architecture Patterns with Python by Harry Percival and Bob Gregory. It's a long read but is well worth checking out.
- ArjanCodes does some great videos on Python programming. This video on the repository pattern helped me redesign a particularly awkward Python script that was performing operations on S3 buckets.