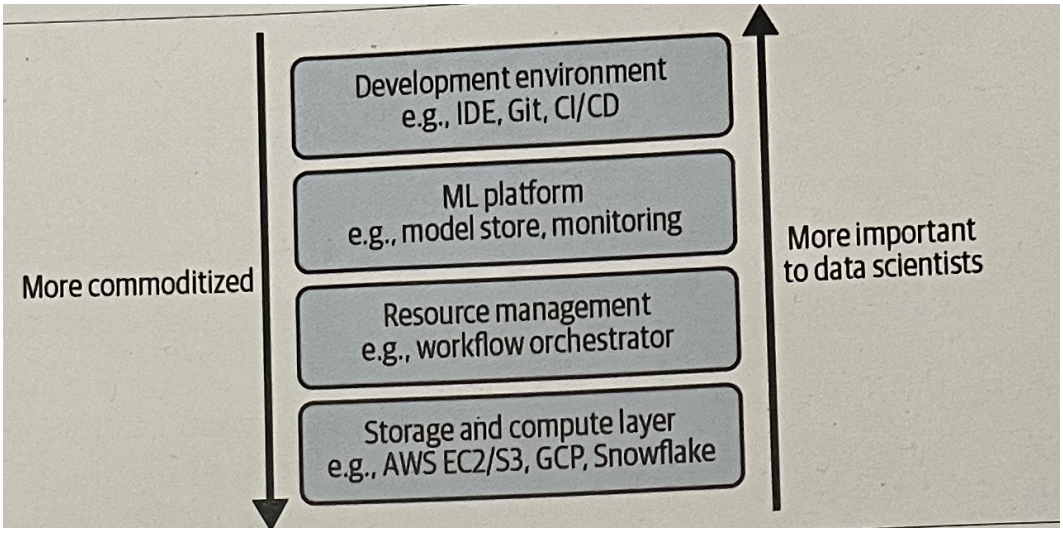
Machine Learning Systems

Created	@April 22, 2023 4:41 PM
Tags	

Chapter 10: Infrastructure and Tooling for MLOps

The book categorize ML infrastructure as 4 layers

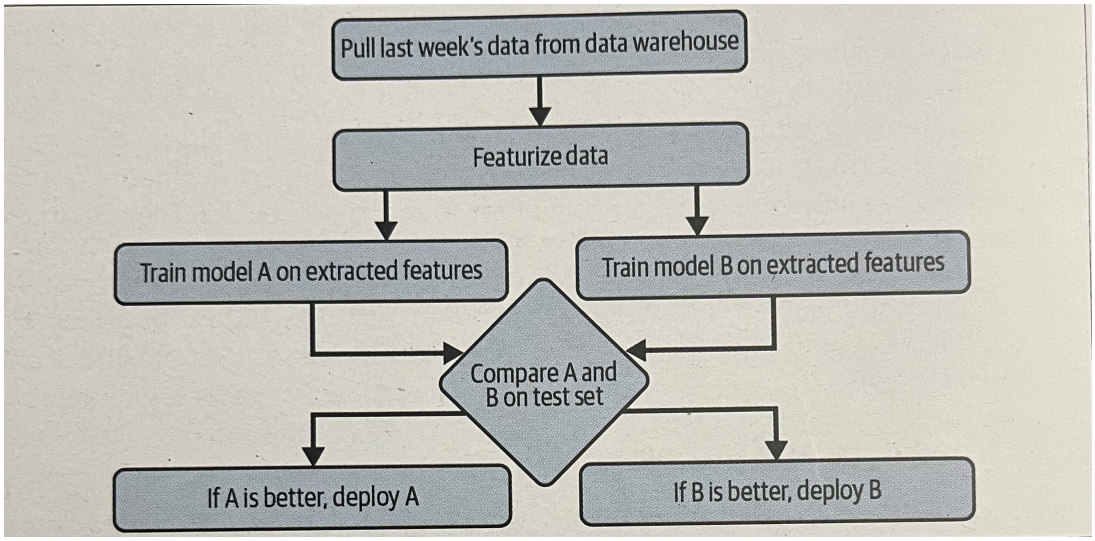


The four layers

- Storage and compute The storage layer where data is collected and stored, providing the compute needed to run your ML workloads such as training a model. computing features, generating features
    - storage: can be hard drive disk in private data center, or cloud like Amazon S3, Snowflake
    - compute: a single CPU/GPU (for cloud, example includes AWS EC2 and GCP); or a computer cluster like K8S
        - Spark and Ray use “job” as their unit, Kubernetes uses “pod”.
        - while you can have multiple containers in a pod, you can’t independently start/stop different containers in the same pod
- Resource management: Resource management comprises tools to schedule and orchestrate your workloads to make the most out of your available

computing resources. Airflow, Kuberflow, Metaflow

- cron: scheduling repetitive works
- scheduler: cron programs that takes DAG (direct acyclic graph)
  - example: pull data, extract feature, train 2 models, and select the better one



- ML platform: tools to aid the dev of ML applications such as model stores, feature stores, and monitoring tools
- dev environment: where code is written and experiments are run
  - need to have CI/CD pipeline setup, tools like GitHub Actions and CircleCI
- production environment
  - docker: to keep production environment the same
    - dockerfiles —build → docker images -run-time→ docker containers
  - docker compose: orchestrate different containers on a single host
  - Kubernetes (K8S): dynamically manage containers, allows resource sharing and easy communication
    More details: https://www.jeremyjordan.me/kubernetes/