Infrastructure and Tooling for MLOps

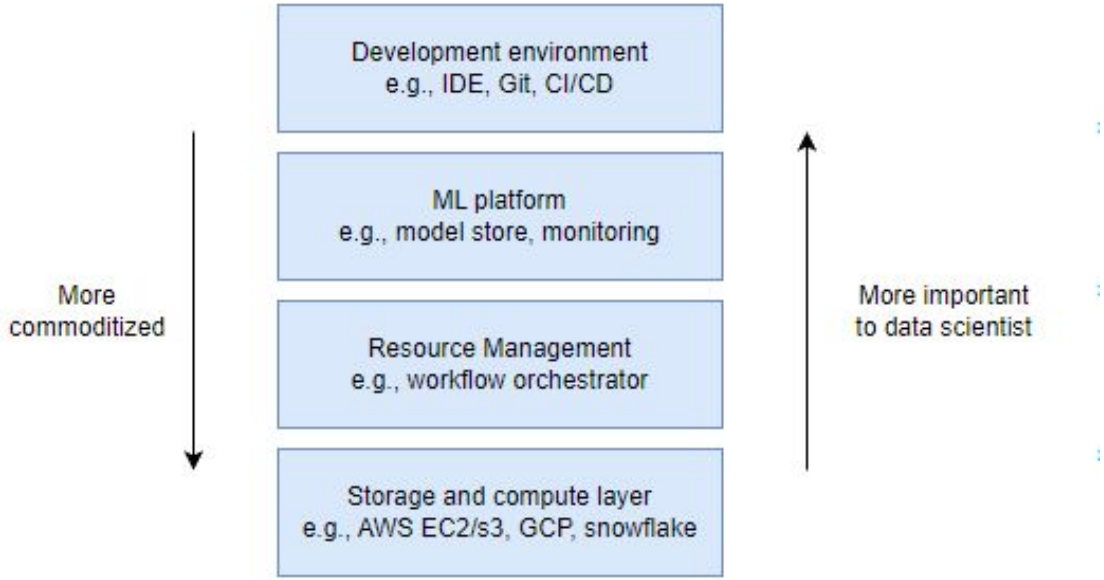
04/12/2023

Reasonable Scale Companies

- End of the spectrum companies
  - Unique requirement
- **Middle of the spectrum companies**
  - Common applications

What Infrastructure means

1. Storage and compute
2. Resource management
3. ML platform
4. Development environment



1. Storage and Compute

- Core/compute unit: CPU/GPU/TPU/IPU with short-lived job, instance
- Compute abstraction: e.g., job, pod
- Most common metric for a compute unit operation speed: FLOPS

Public Cloud vs Private Data Centers

Public Cloud:

No up front, elastic, unattended,省时/省钱/省力

Data Centers:

https://dev.37signals.com/our-cloud-spend-in-2022/

https://tech.ahrefs.com/how-ahrefs-saved-us-400m-in-3-years-by-not-going-to-the-cloud-8939dd930af8

Multicloud Strategy

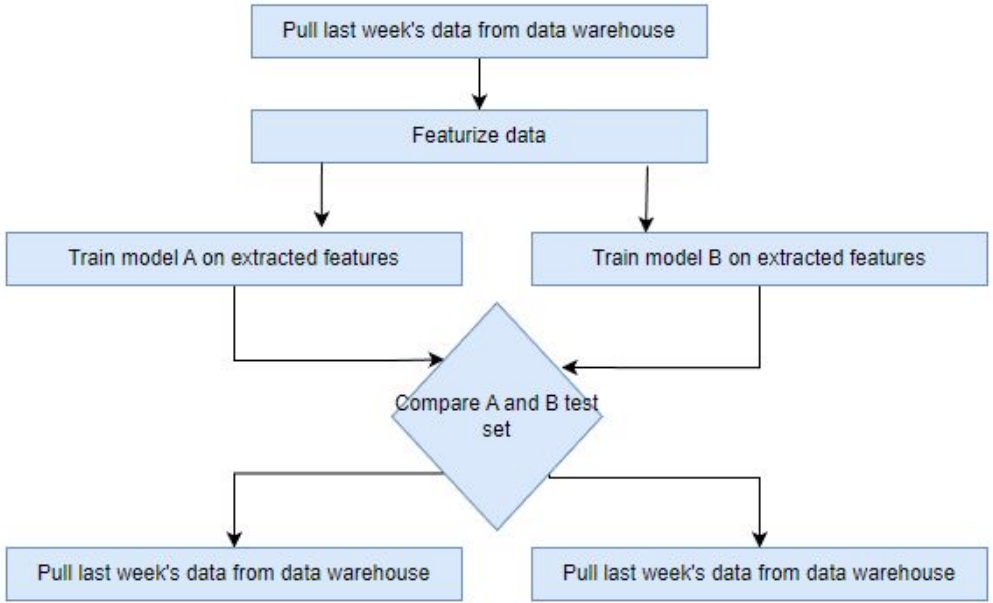
https://www.gartner.com/smarterwithgartner/why-organizations-choose-a-multicloud-strategy

2. Resource Management - Cron, Scheduler, Orchestrator

- Manage compute resources for ML workflows
- Two key characteristics of ML workflows that influence their RM:
  - Repetitiveness
  - Dependencies
- Scheduler: **when** to run jobs and **what** resources are needed, dealing with DAGs, PQs, or User-level quotas
- Orchestrators: **where** to get these resources, dealing with lower-level abstractions like machines, instances, clusters, replications, etc.

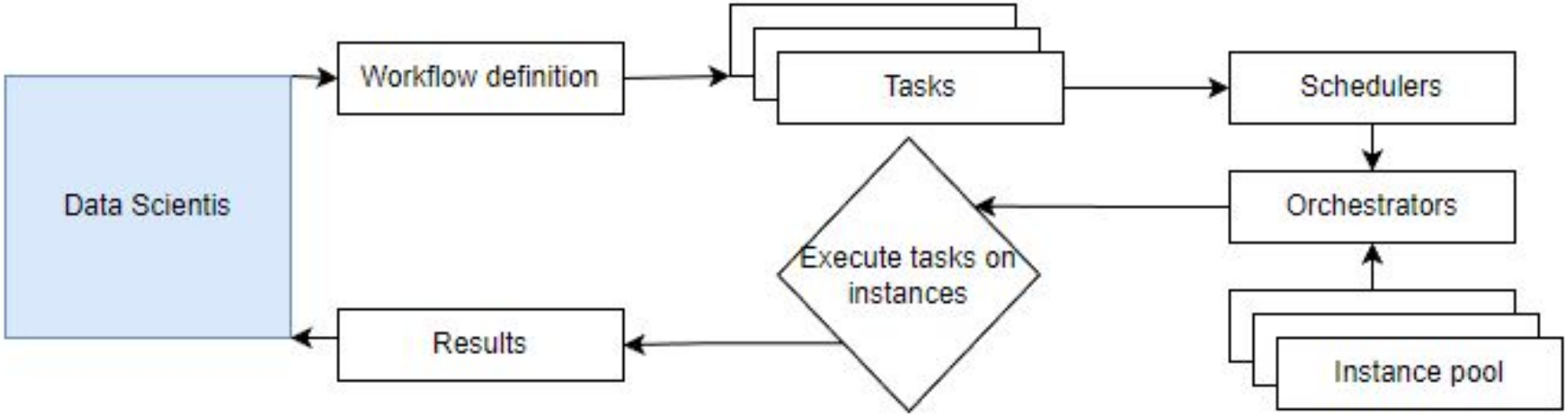
Example

1. Pull last week’s data from data warehouse
2. Extract features from this pulled data
3. Train 2 models, A and B, on the extracted features
4. Compare A and B on the test set
5. Deploy A if A is better; otherwise deploy B



Workflow Management

After a workflow is defined, the tasks in this workflow are scheduled and orchestrated



Most common workflow management tools(Orchestrators)

- Airflow
- Argo Workflow
- Prefect Workflow
- Kubeflow
- Metaflow

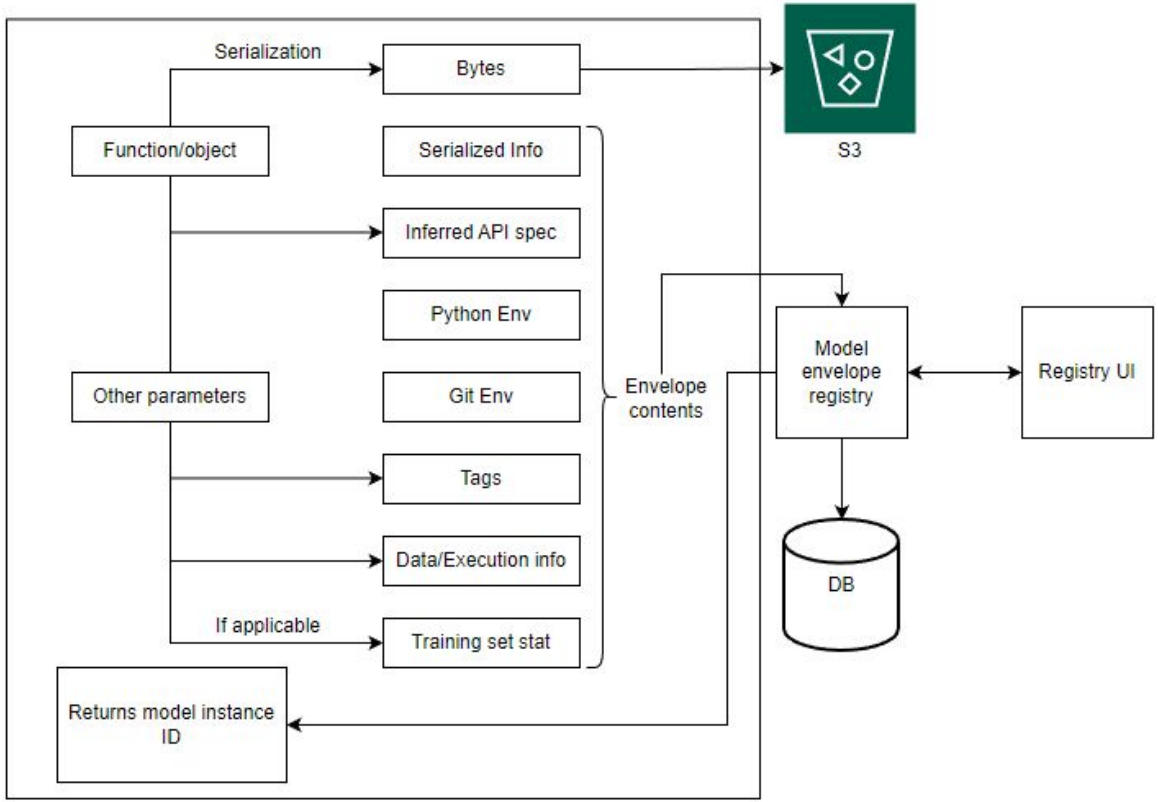
3. ML Platform

Evaluation: cloud provider or DC, open source or managed service

Model Management

- Model Deployment
  - SageMaker, Azure ML, Vertex AI
- Model Store
  - Blob storage, e.g., S3
  - Track more information (model definition, model parameters, featurize and predict functions, dependencies, data, model generation code, experiment artifacts, tags)
- Feature Store
  - Feature management
  - Feature computation
  - Feature consistency

Model management example -Stitch Fix



4. Dev Environment

Local IDE: VS Code, Vim, Jupyter Notebooks (+ Papermill, commuter, nbdev)

Cloud IDE: AWS Cloud9/SageMaker Studio, Google Colab, Github Codespaces

Dev to Prod

- Autoscaling
- HA
- Dockerfile -> Registries

References