Building a Reproducible Model Workflow Cont.

Final Pipeline, Release and Deploy

UFRN UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

DCA

Table of Contents

01

Decision Tree
Introduction, Mathematical Foundations

02

Evaluation Metrics
Best practices, threshold and ranking metrics

An inference pipeline is an ML pipeline that contains everything that needs to run in production at inference time: a pre-processing step that transforms the data input to the data expected by the model, and then the model

03

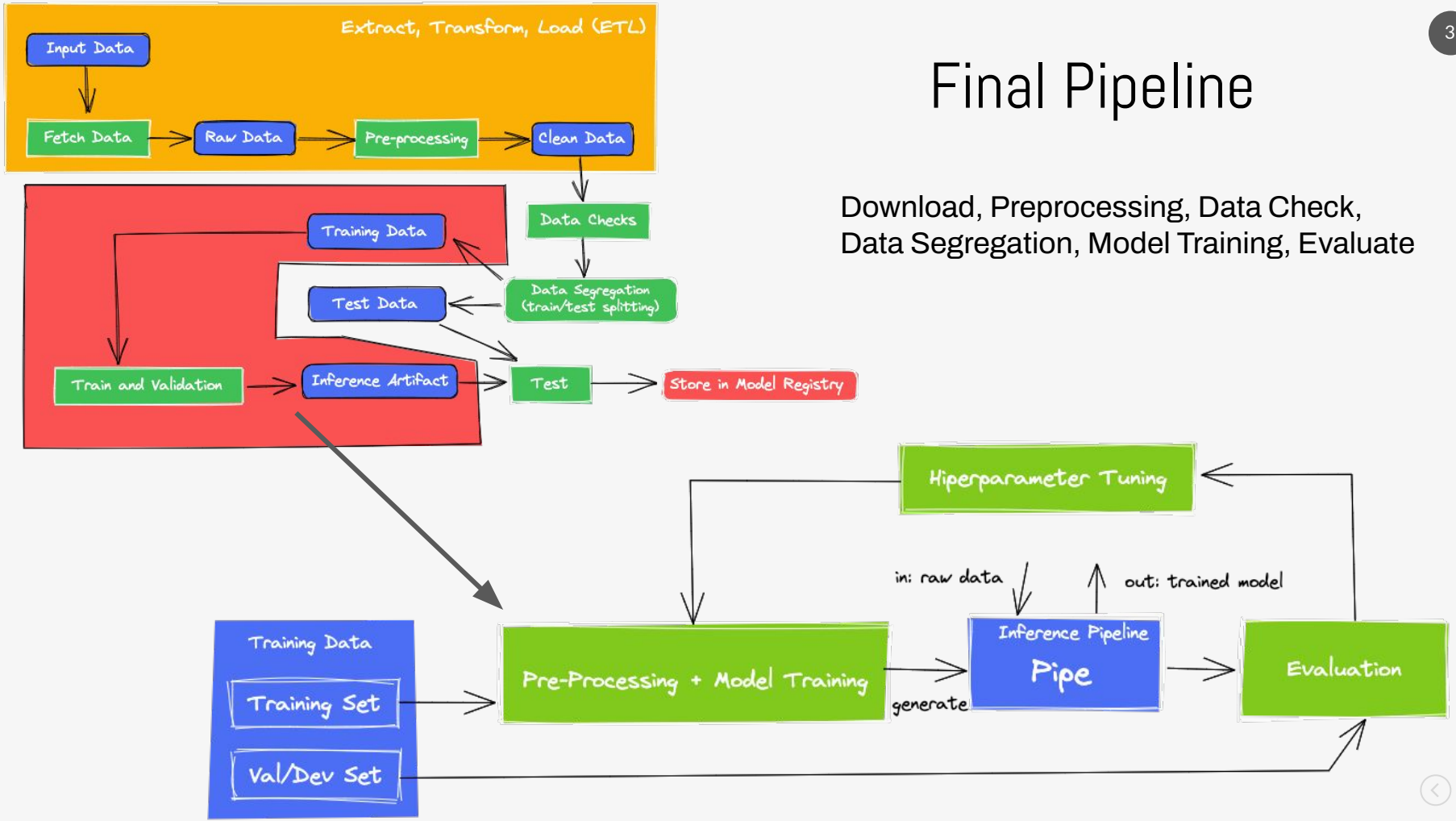
Implementing Pipelines
From MLOps 0 to 1

04

Test Evaluation

Final Pipeline

Download, Preprocessing, Data Check,
Data Segregation, Model Training, Evaluate

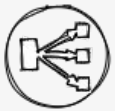


ML Pipeline in MLFlow

A project that calls other projects (the components)

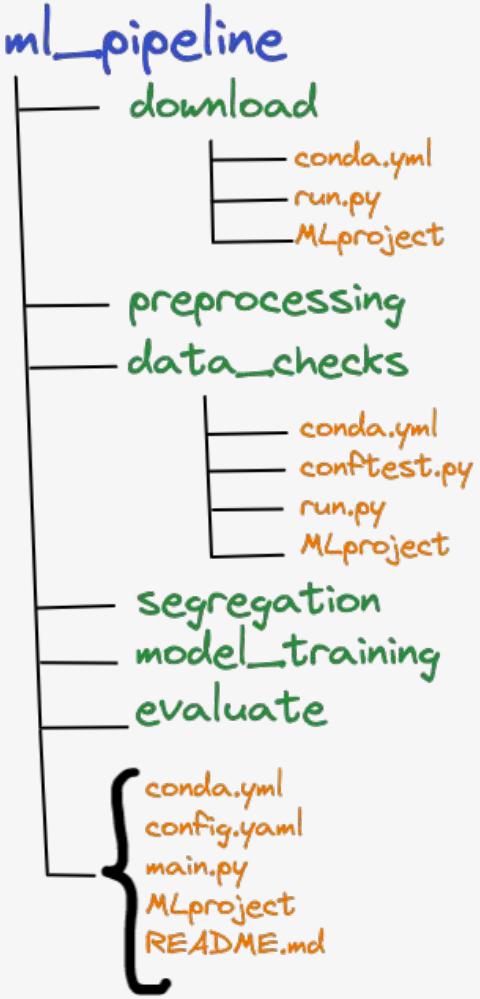
One repository

For simplicity



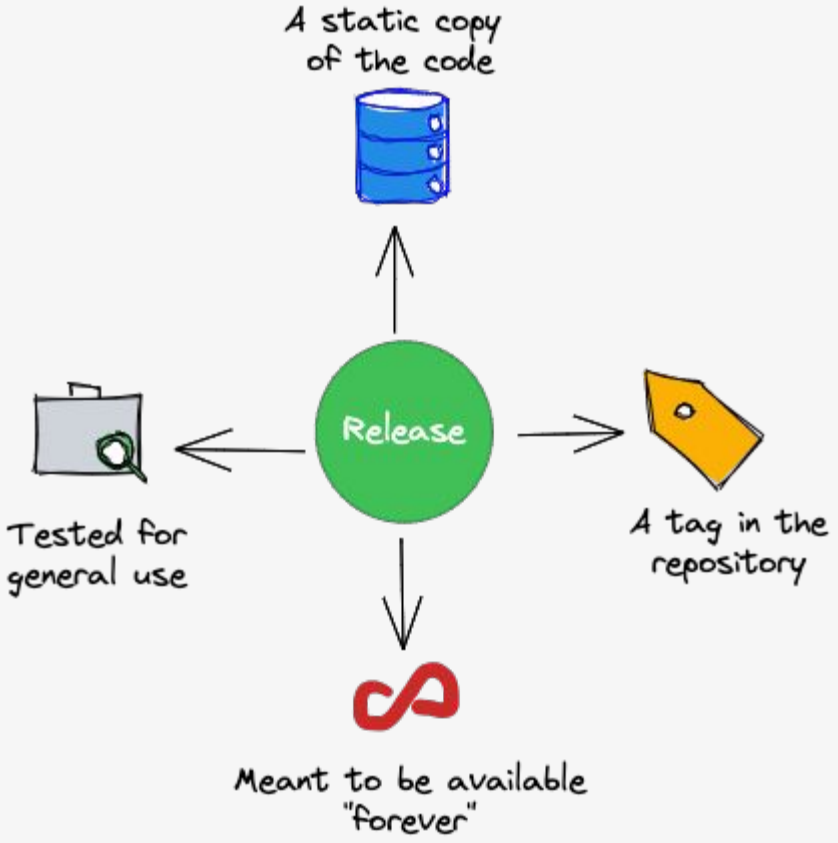
Another option

When your operations scale up, you will want one or more repos for the components and one or more repos for each project



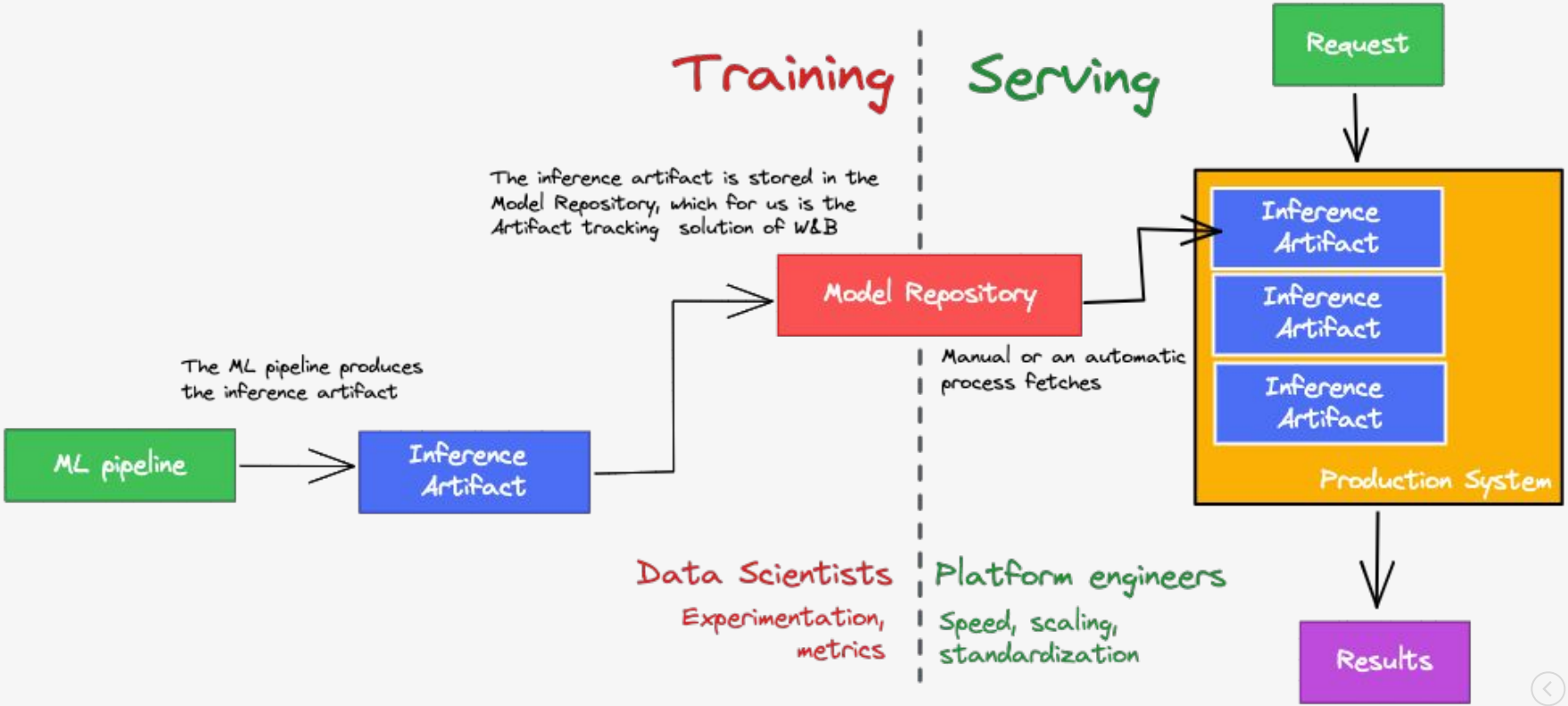
What is a release?

A release is a static copy of the code that reflects the state of the code at a particular point in time. It has a version attached to it, and a tag. The tag can be used to restore the repository (or a local copy of the code in the repository) to the state it was when the release was cut.



Semantic Versioning

Deployment


Training
Serving
Request
The inference artifact is stored in the Model Repository, which for us is the Artifact tracking solution of W&B
Model Repository
Inference Artifact
Inference Artifact
Inference Artifact
Production System
Manual or an automatic process fetches
The ML pipeline produces the inference artifact
ML pipeline
Inference Artifact
Data Scientists
Experimentation, metrics
Platform engineers
Speed, scaling, standardization
Results

Using a model in production: Online vs Offline Inference