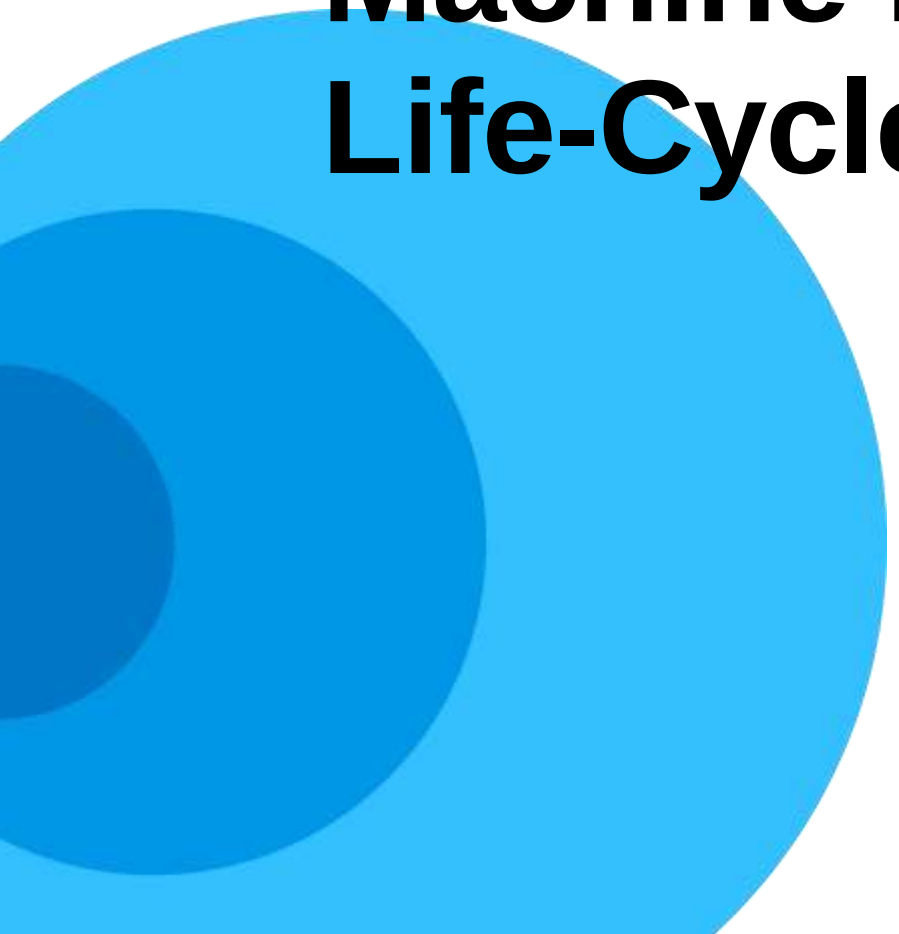


Machine Learning Platform Life-Cycle Management

Hope(Xinwei) Wang
Software Engineer at Intuit

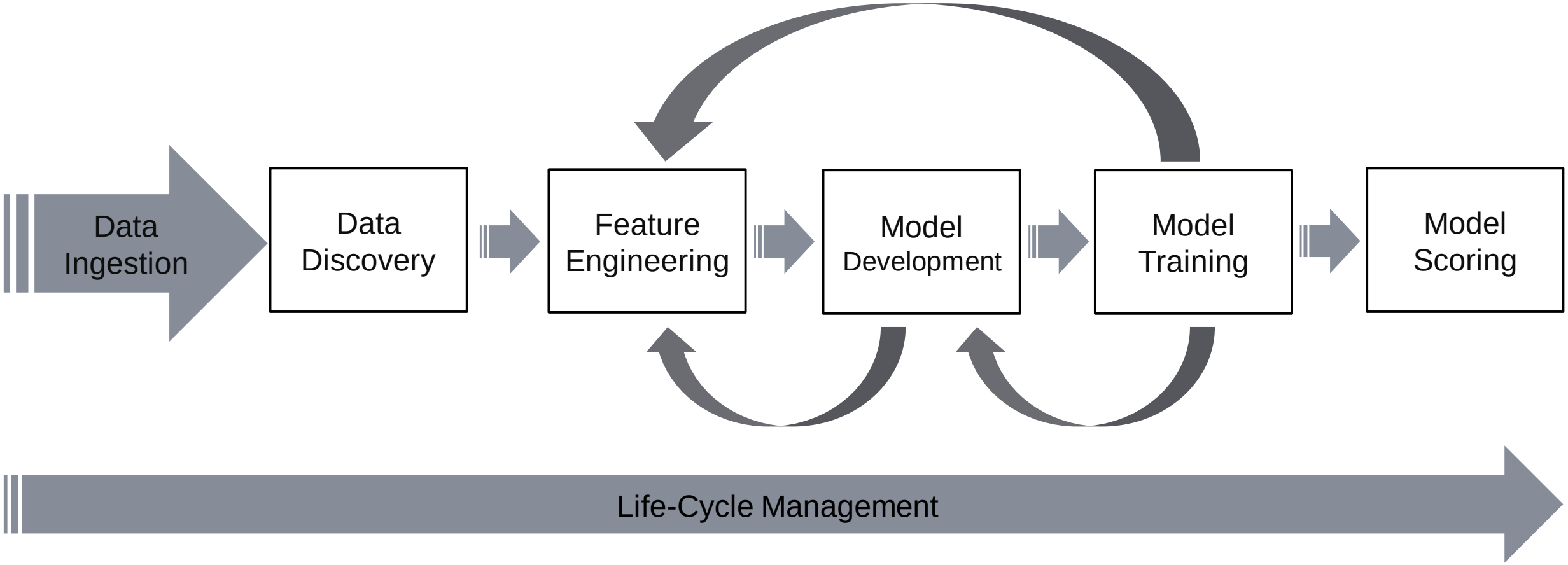
OVERVIEW

- What is a machine learning platform?
- What is the ML platform lifecycle?
- Why ML platform lifecycle management?
- Artifacts and their associations
- Use cases at Intuit

What is a Machine Learning Platform?

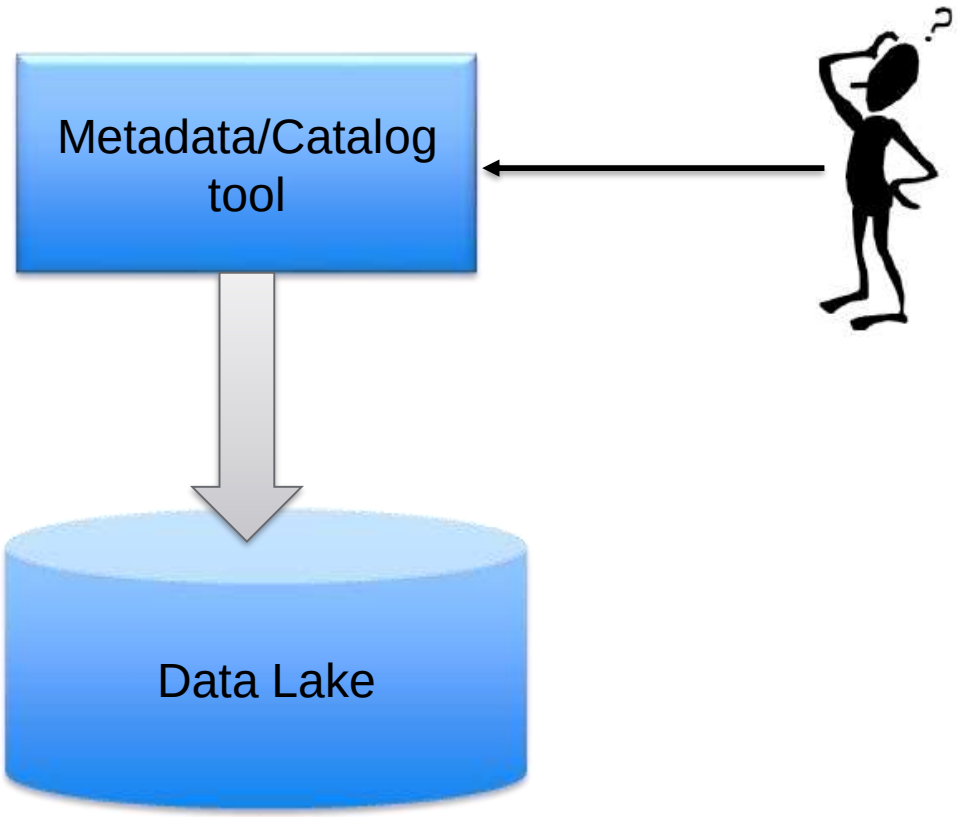
- Manages the entire lifecycle of an ML model
- Includes automating and accelerating the delivery of ML applications

ML Platform Life-Cycle



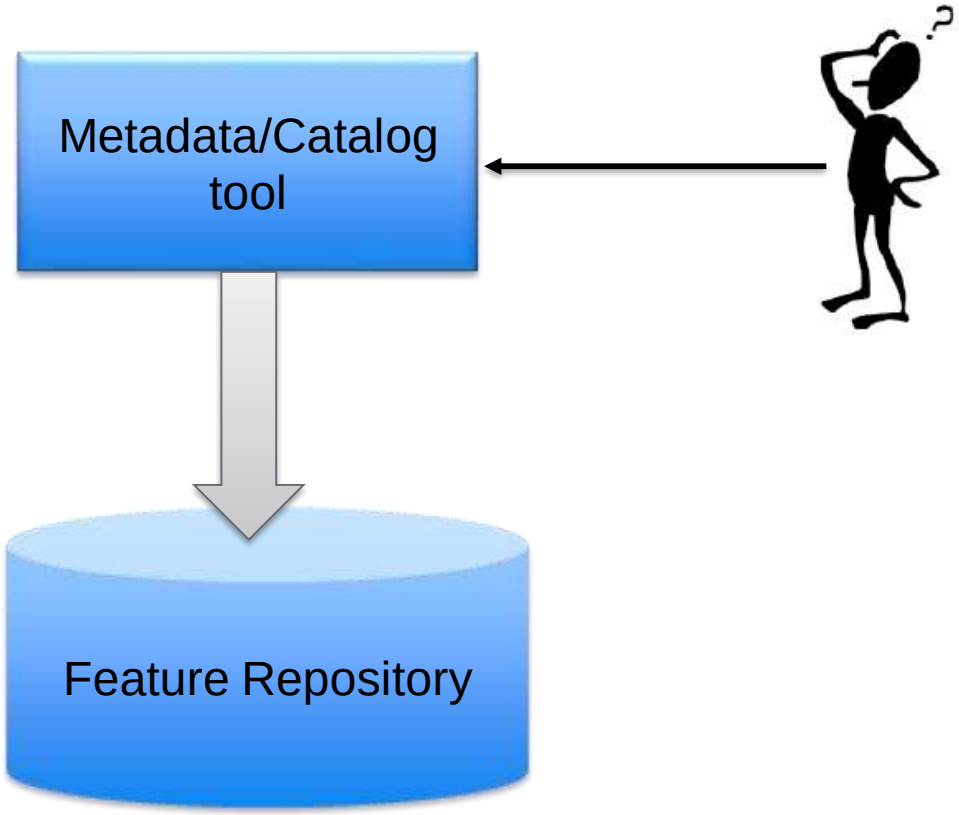
Data Discovery

- Metadata/catalog tool
- Accessible data source
  (Raw attributes & Data lineage)



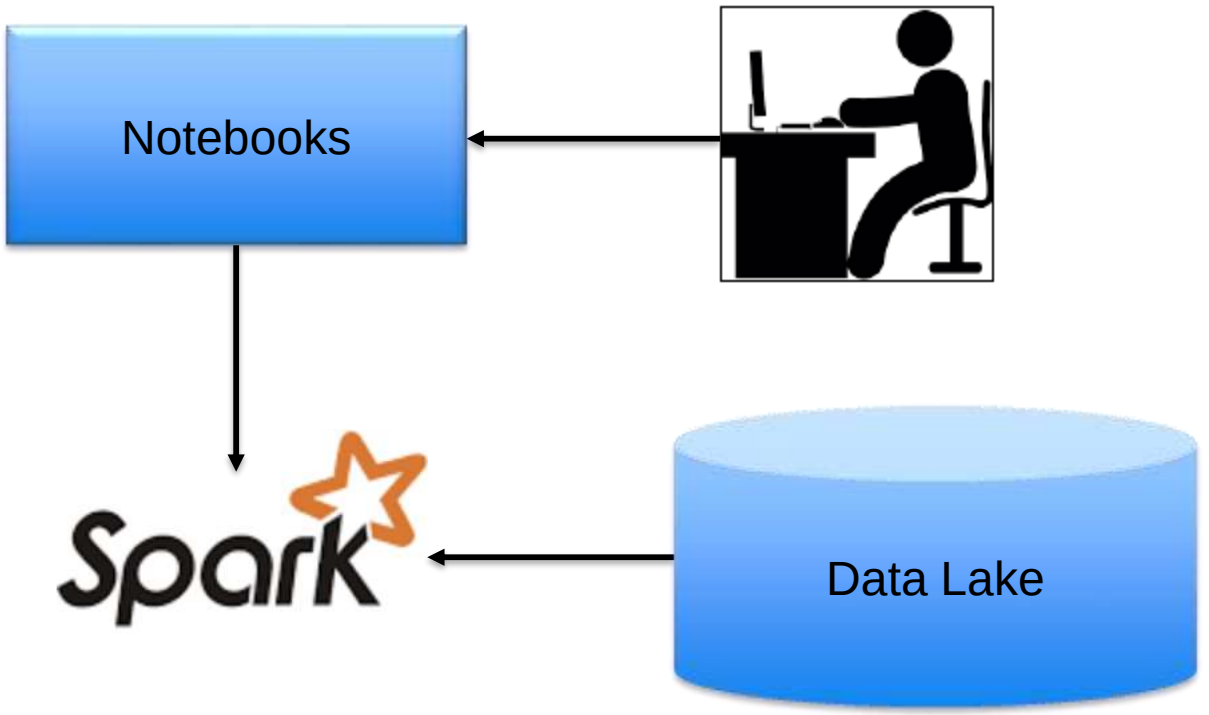
Feature Engineering

- Output : features
- Reproducible
- Reusable



Model Development

- Collaborative environment
- Access data lake



Model Training

Support ability for:

- Being triggered either manually/via automation
- Creation and management of training sets
- Re-training
- Optimizing *hyper parameter tuning* through parallelization of model training execution

Model Scoring

- Support online/offline(depend on use cases)
- Ability to be triggered either manually/via automation

Big Mess!

- No central artifact management solution
- Hard to reuse existing features/data/algorithms/toolings
- Inability to scale for large datasets
- Lack of automation/orchestration across the ML life-cycle
- Lack of rigor/discipline in the ML development life-cycle
- **Slow down delivery of Machine Learning applications**

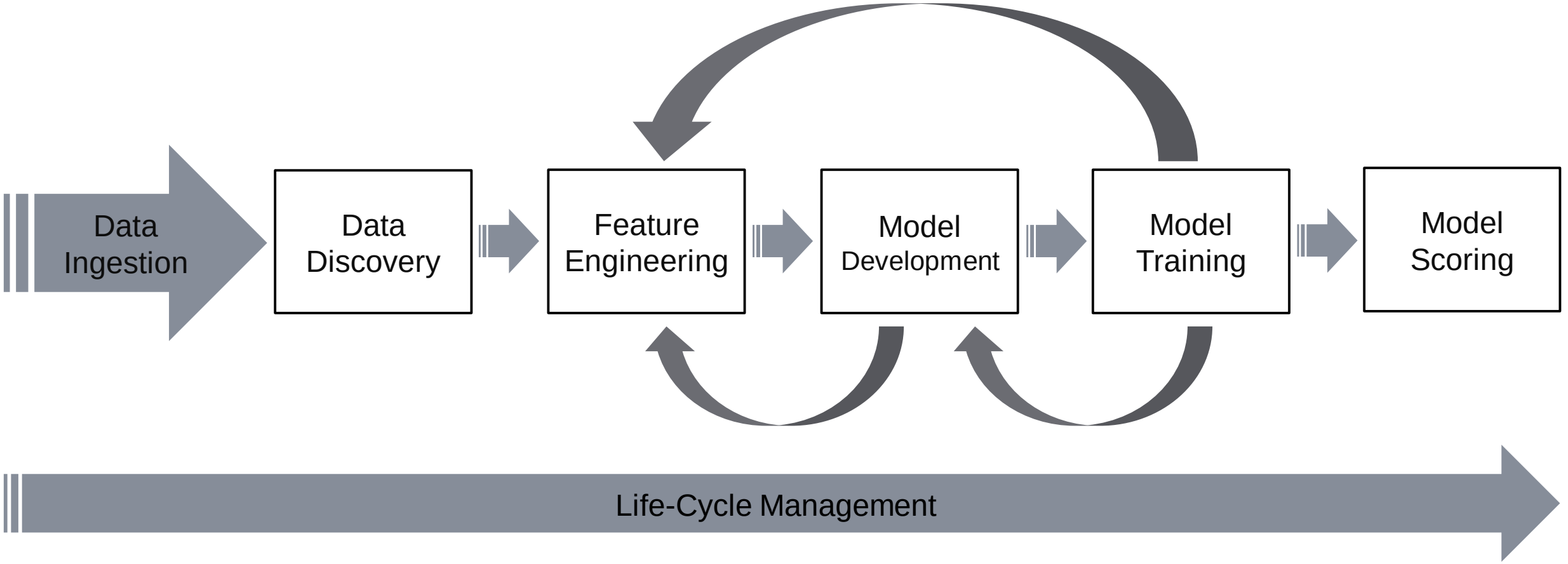
Ideal Status

- Optimizing data scientists' engineering process
- Tie ML components together into a cohesive platform, support the life-cycle of ML artifacts end-to-end
- Increase efficiency of delivering ML predictions at scale

Artifact Management

- **Data Artifacts**
  - Features
  - Training sets
- **Model Artifacts**
  - Model code
  - Trained models
  - Performance metrics
  - Hyper parameter values
- **Environment Artifacts**
  - Languages & language versions
  - Packages & Package versions

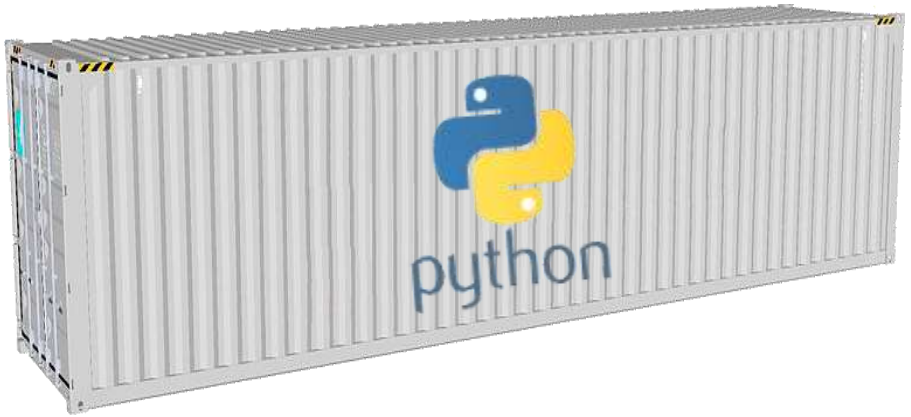
Machine Learning Platform Life-Cycle Management



CONTAINERIZATION!

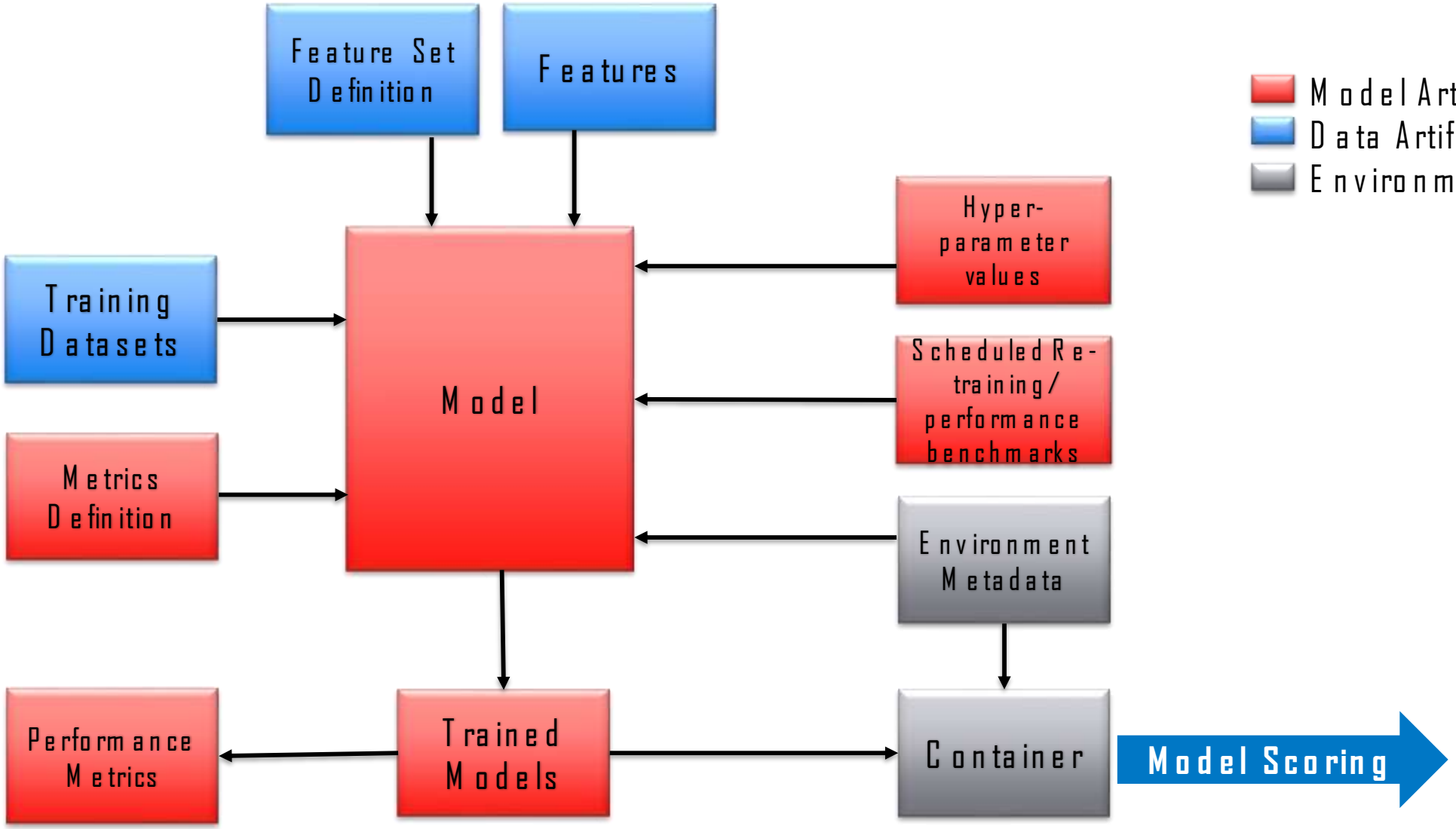
Environment in Container

Benefit Of Containerization



- **Flexibility :** Model has specific environment
- **Consistency :** Model has same behavior throughout the life-cycle

Machine Learning Platform Life-Cycle Management



Model Code

- Developed in notebooks
- Multiple versions
- Each version associate with an externalized environment artifact

Environment Artifacts

- Environment must be consistent for development, training, scoring
- Externalized as metadata
- Model/Execution environments constructed from metadata and deployed into containers (Docker, Yarn, Conda, etc.)

Examples of containers/virtual environments

- Tailored to the environment (built based on externalized environment metadata)
- Used for model development, training, execution

Container/virtual environment	Usage
Docker container	• Model development • Model training • Online scoring
Yarn container	On Spark cluster • Distributed training • Batch offline scoring
Conda environment	Model development

Features

- Used as data input of the model
- Stored in discoverable feature repository
- Metadata defines the model specific feature-sets

Trained Models

- Serialized, weighted model files
- Associate with a version of model code and training set

Training Sets

- Datasets used to train, validate and test the model
- Associated to a trained model

Feature Set Definitions

- Define what feature sets this model requires

Hyper Parameter Values

- Set up values before learning process
- Model specific

Metric Definition

- Defines the metrics to collect and thresholds to evaluate models against.

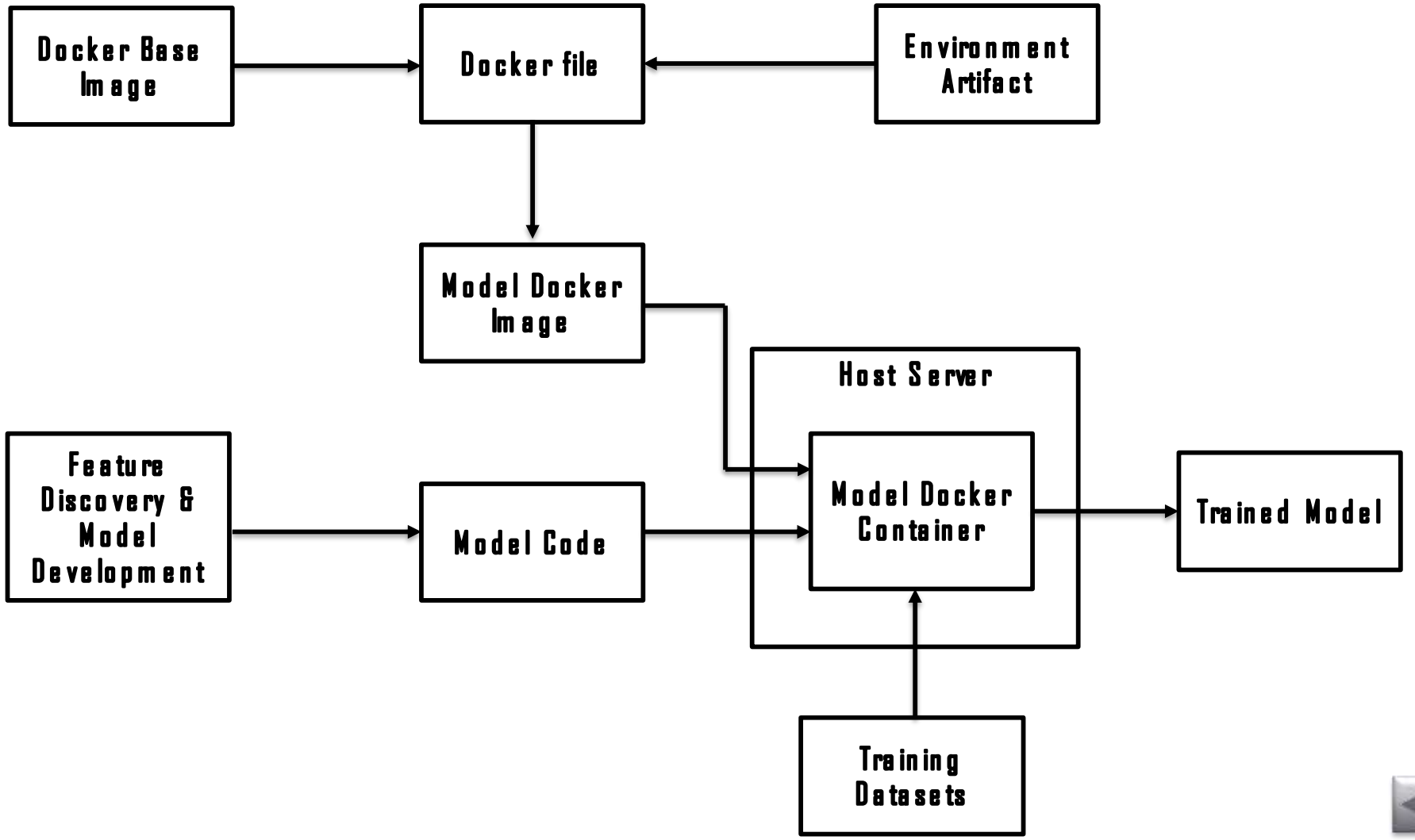
Performance Metrics

- Metrics to evaluate model effectiveness
- Model metrics including: ROC curve, confusion matrix, F1 score, precision, recall, etc.

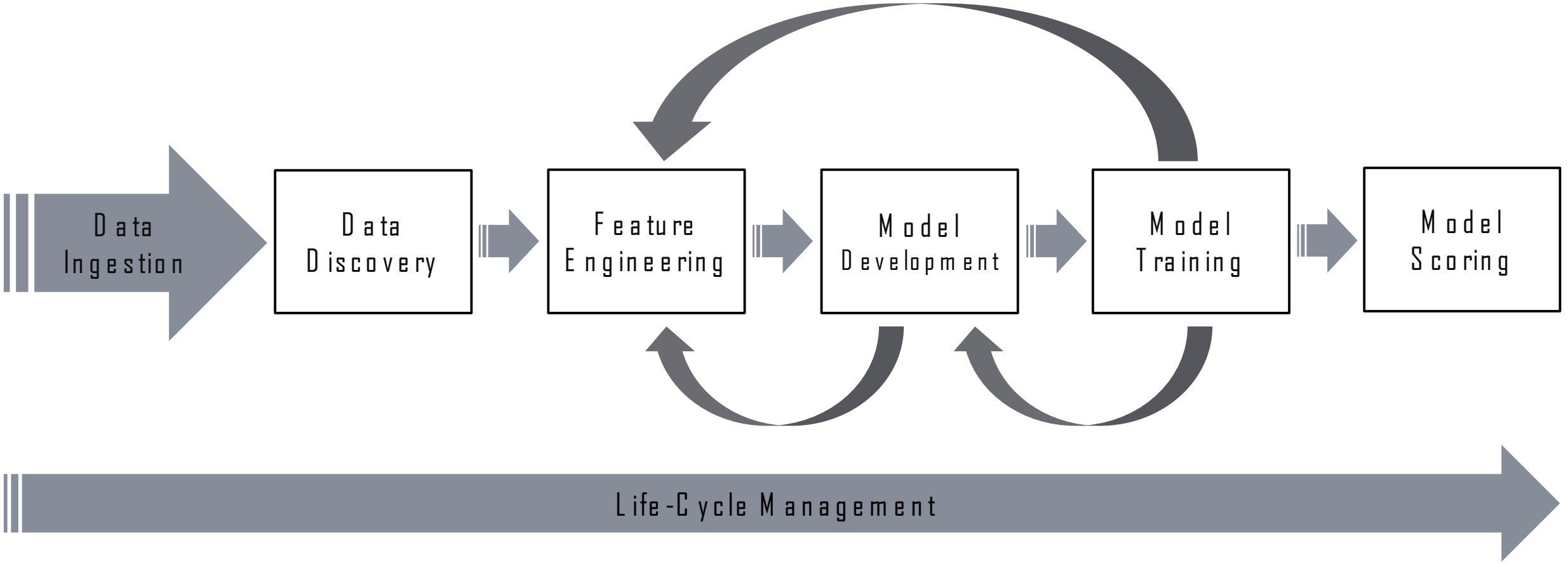
Scheduled Re-training & performance benchmarks

- To automate the re-training and deployment of updated models
- Model specific

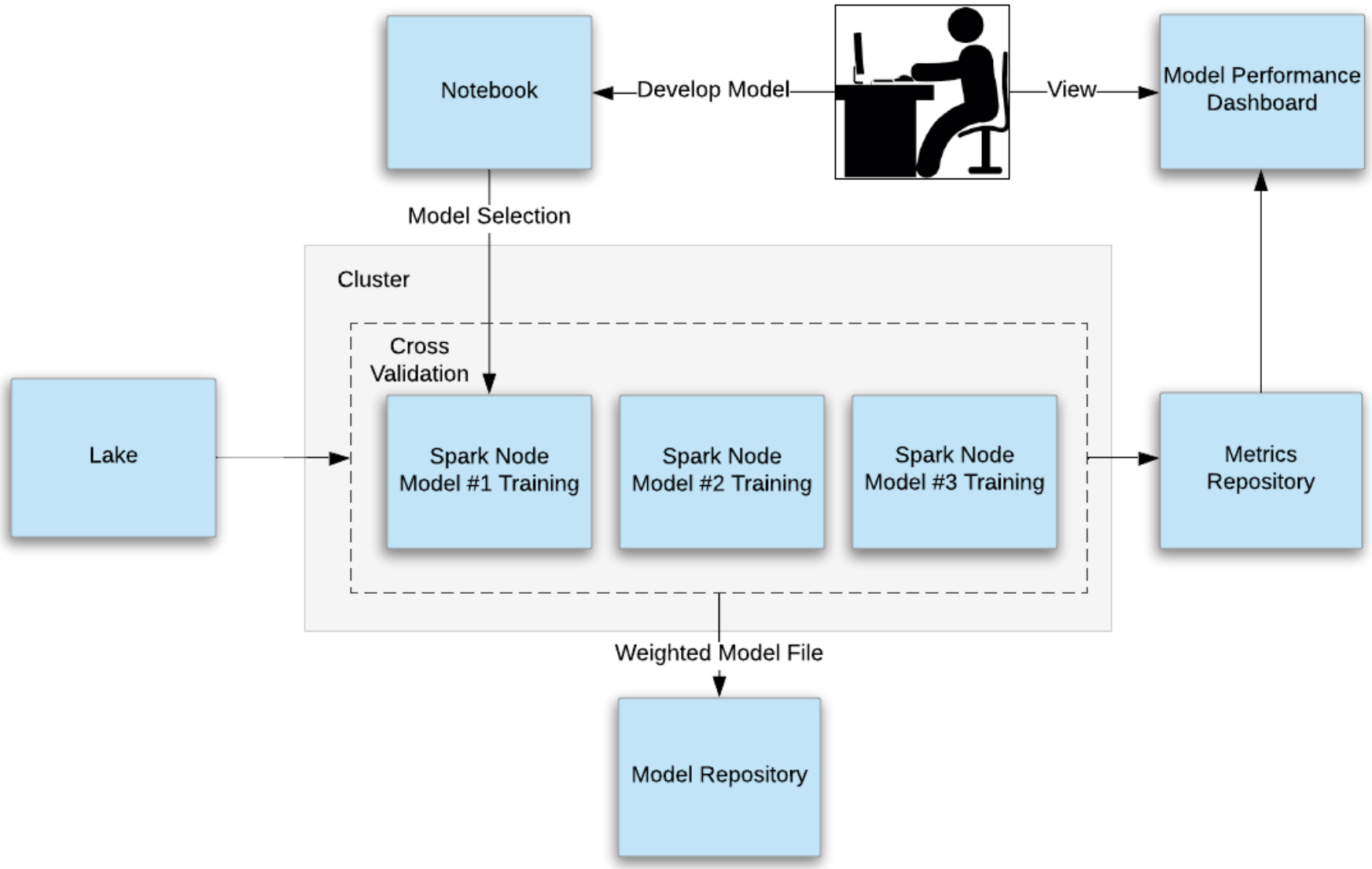
Model Training in Docker Container



Machine Learning Platform Life-Cycle Management



Model Development & Training & Tuning



Self-help Service in Quickbooks

intuit quickbooks

Get Contextual help

Personalization Service

Clickstream Data

Feature Service

Score Data

Batch Source

Kafka

Spark Streaming

Online Features

Online Store

Online Scoring Service

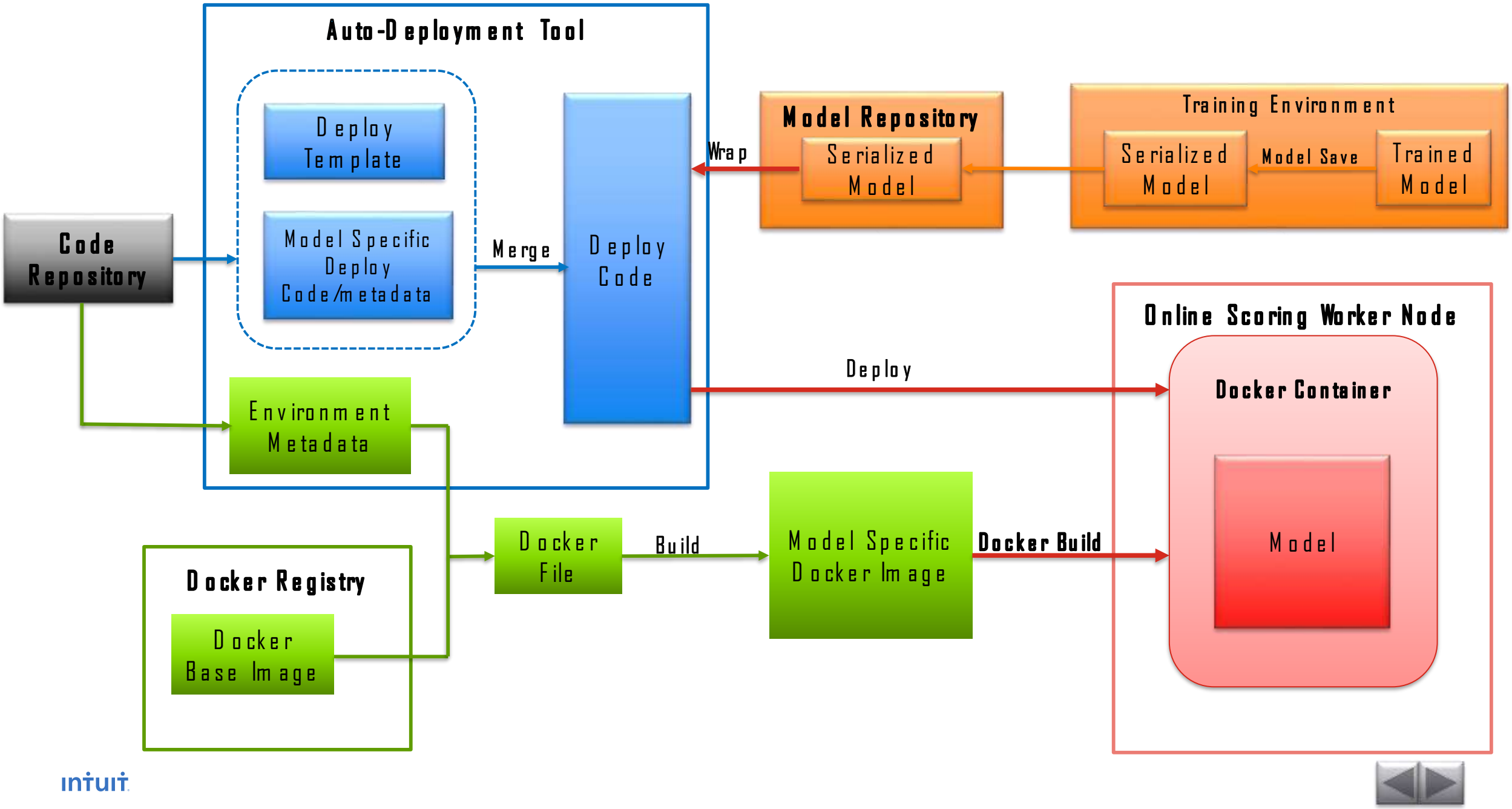
Offline Features

Lake

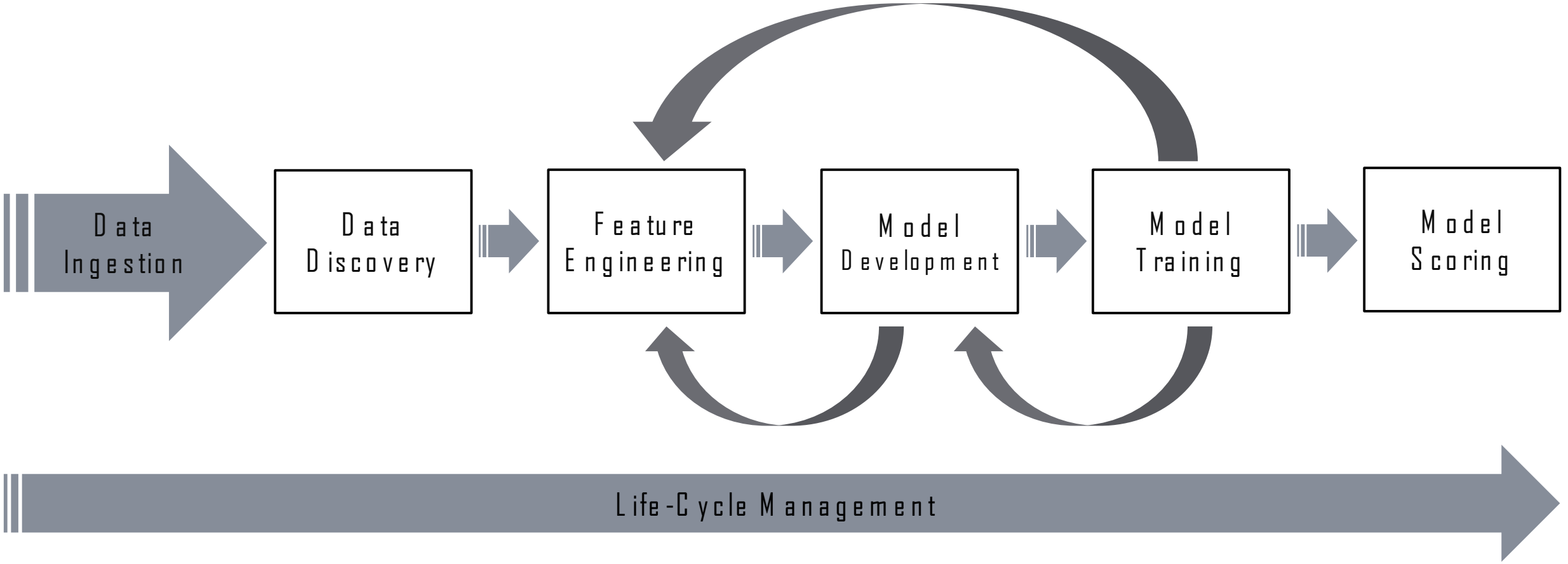
Offline Spark Job

Feature Repository

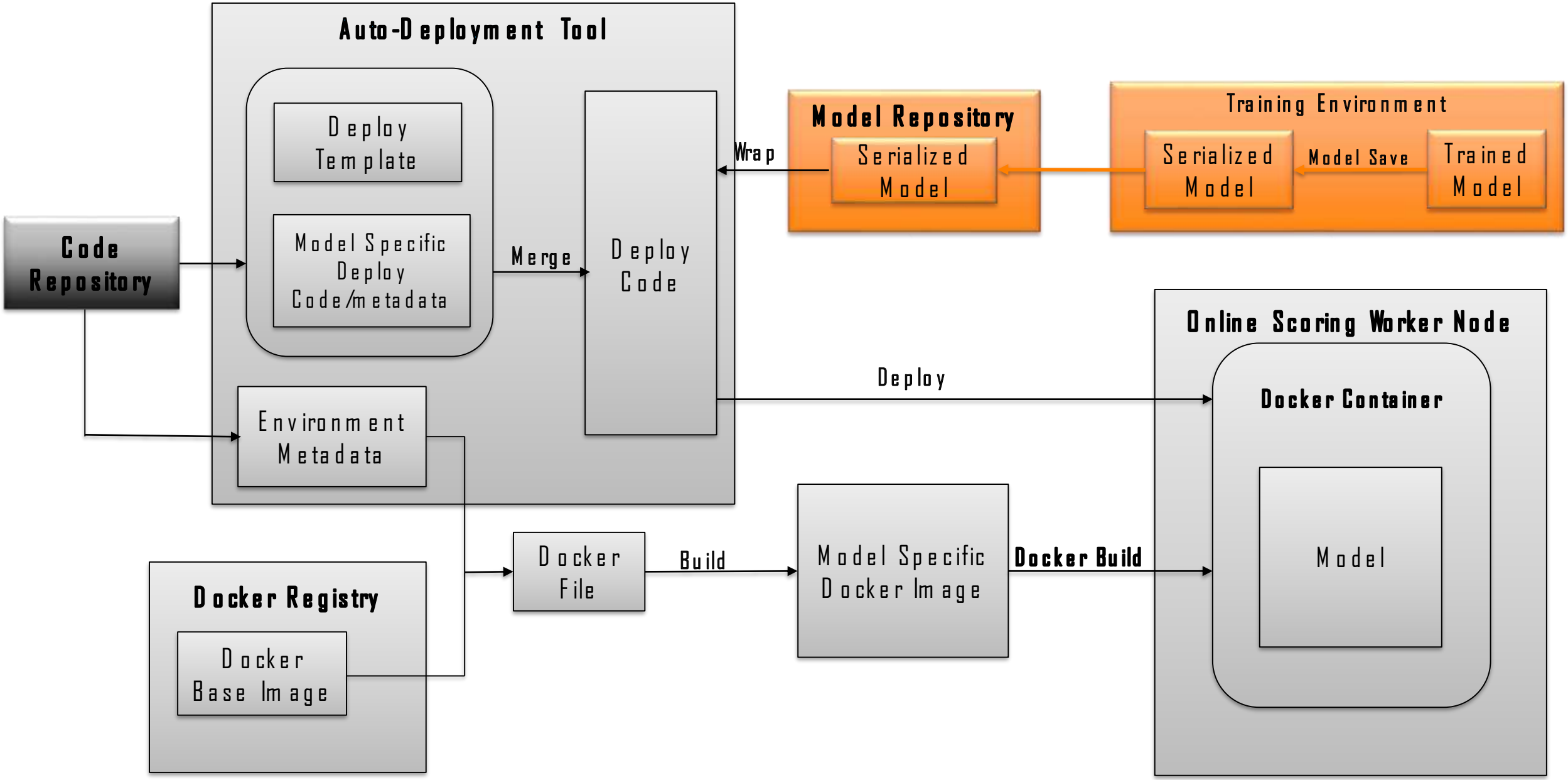
Example: Online Scoring Service Deployment Diagram



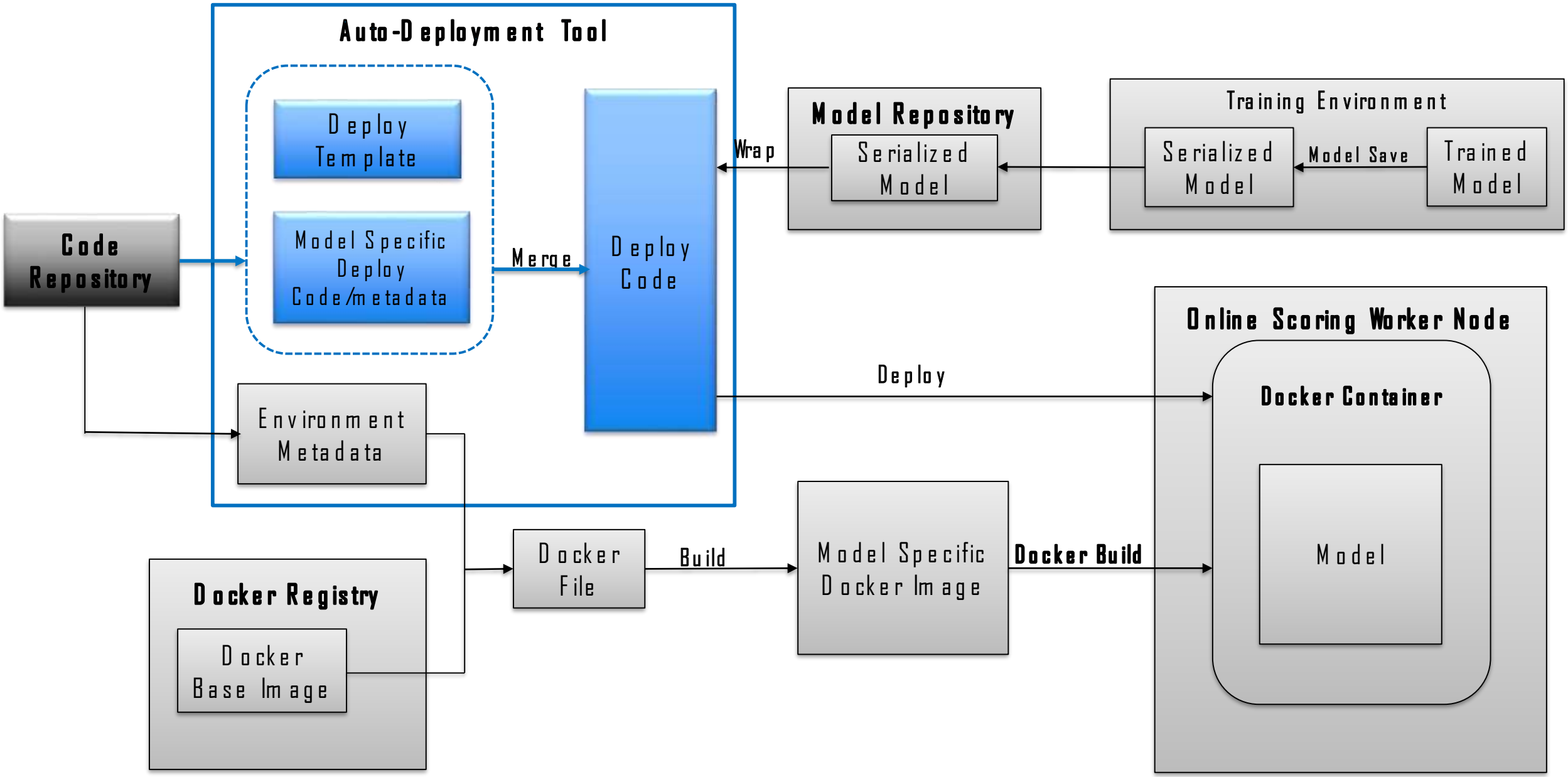
Machine Learning Platform Life-Cycle Management



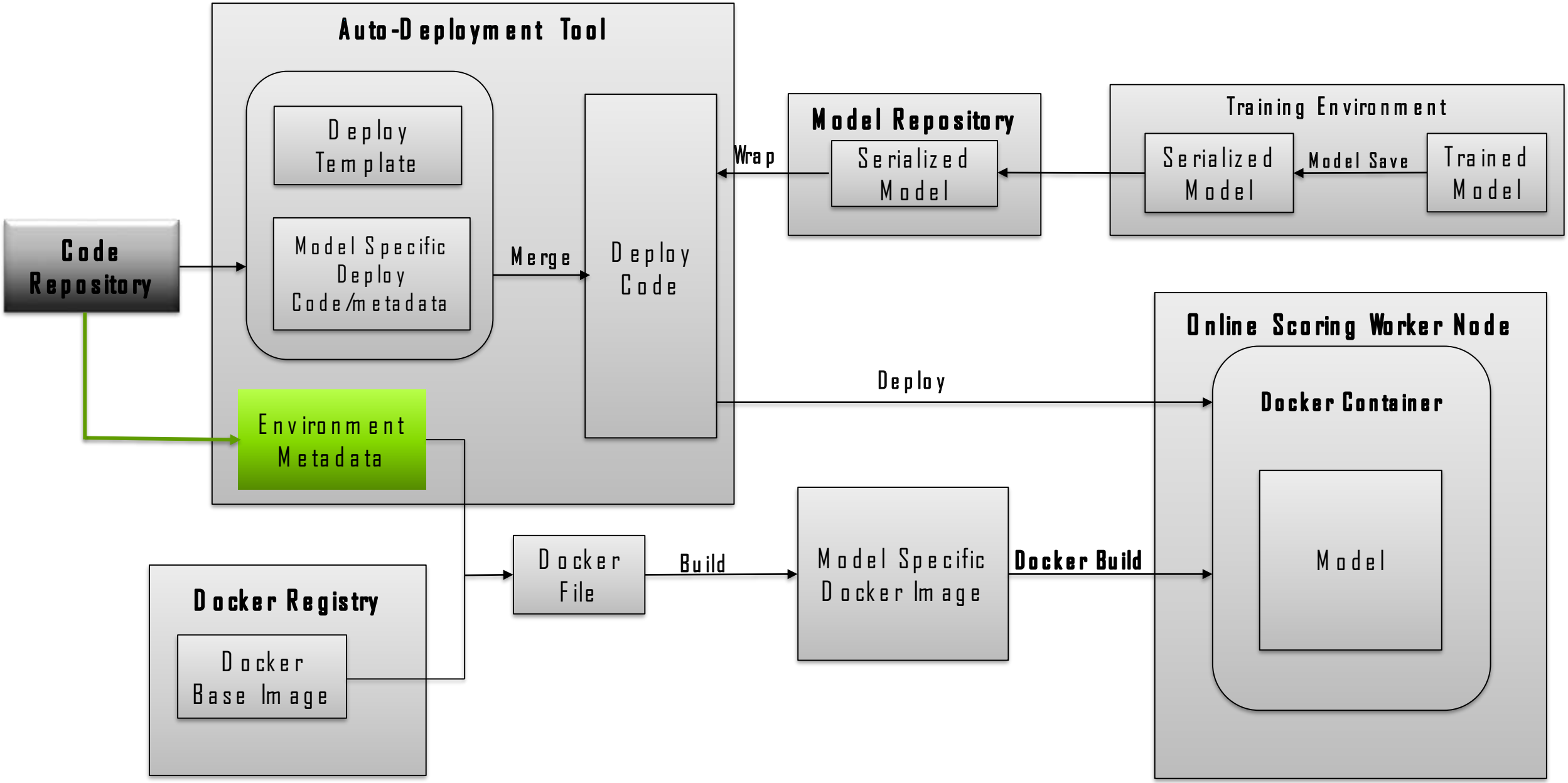
Example: Online Scoring Service Deployment Diagram



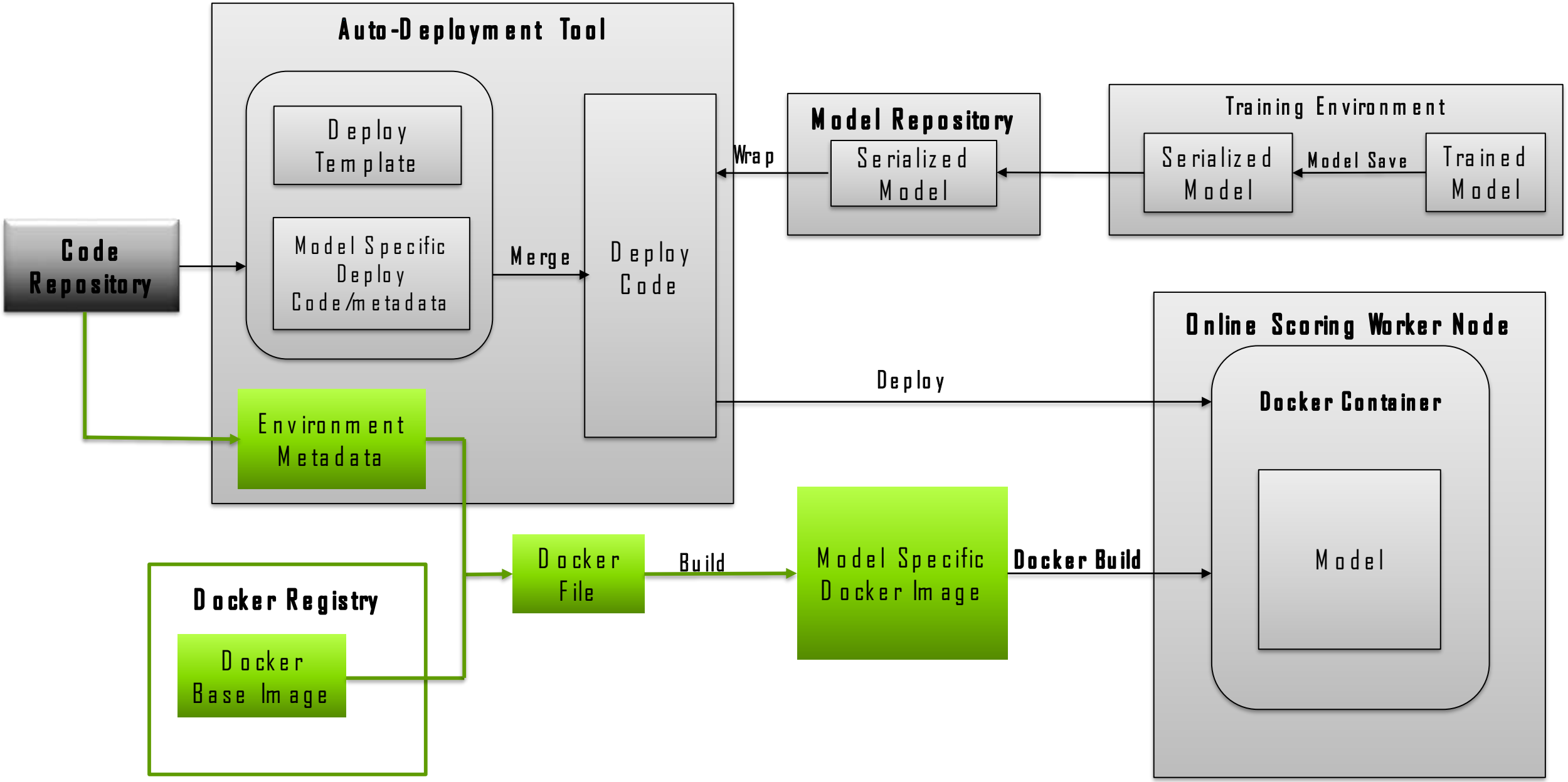
Example: Online Scoring Service Deployment Diagram



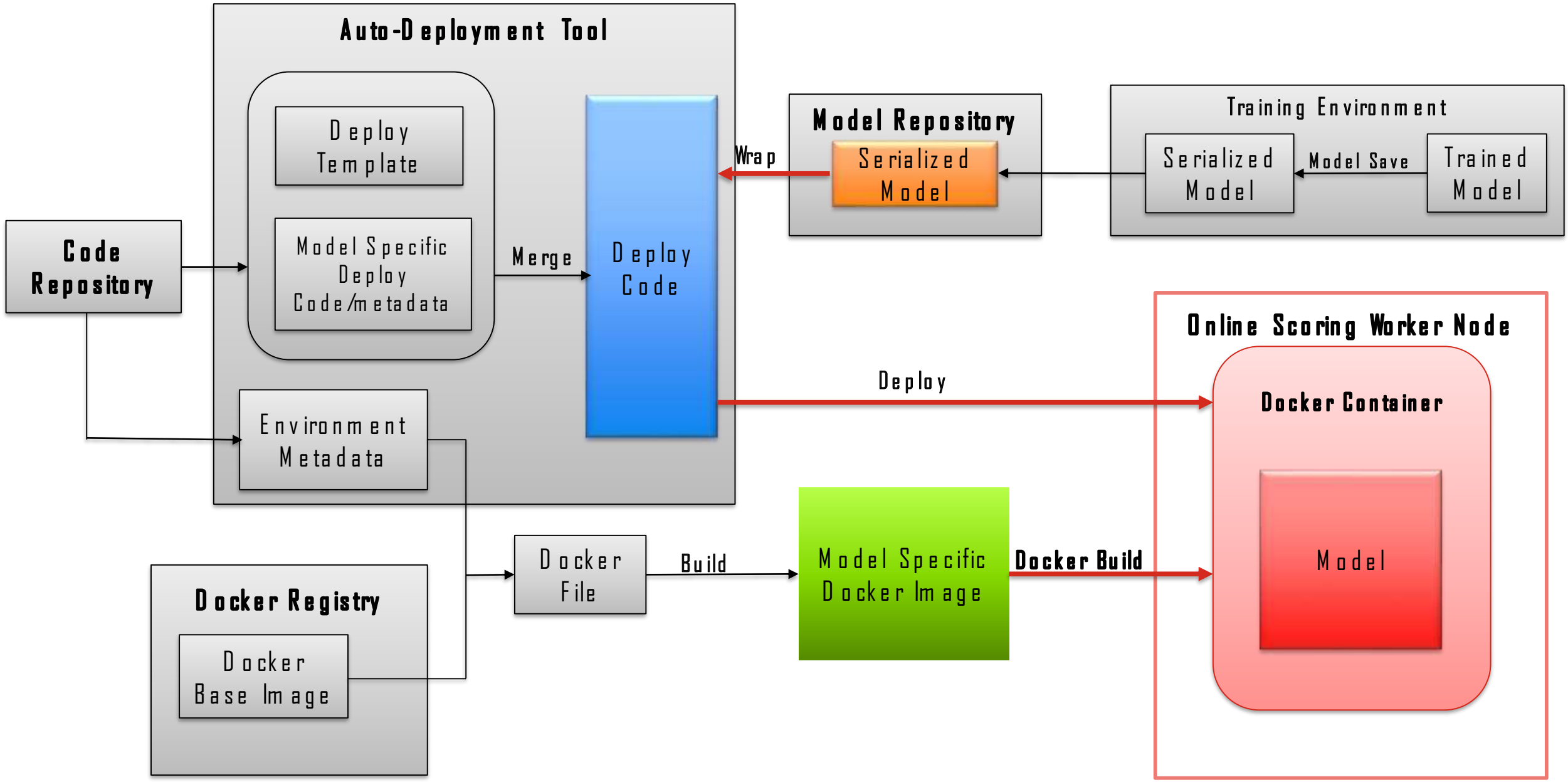
Example: Online Scoring Service Deployment Diagram



Example: Online Scoring Service Deployment Diagram



Example: Online Scoring Service Deployment Diagram



CONTAINERIZATION!

Environment in Container

Thank you

Hope (Xinwei) Wang

Email: xinweiwang3@gmail.com

Linkedin: https://www.linkedin.com/in/xinweiwanglinkedin/

Twitter: @HopeXinwei