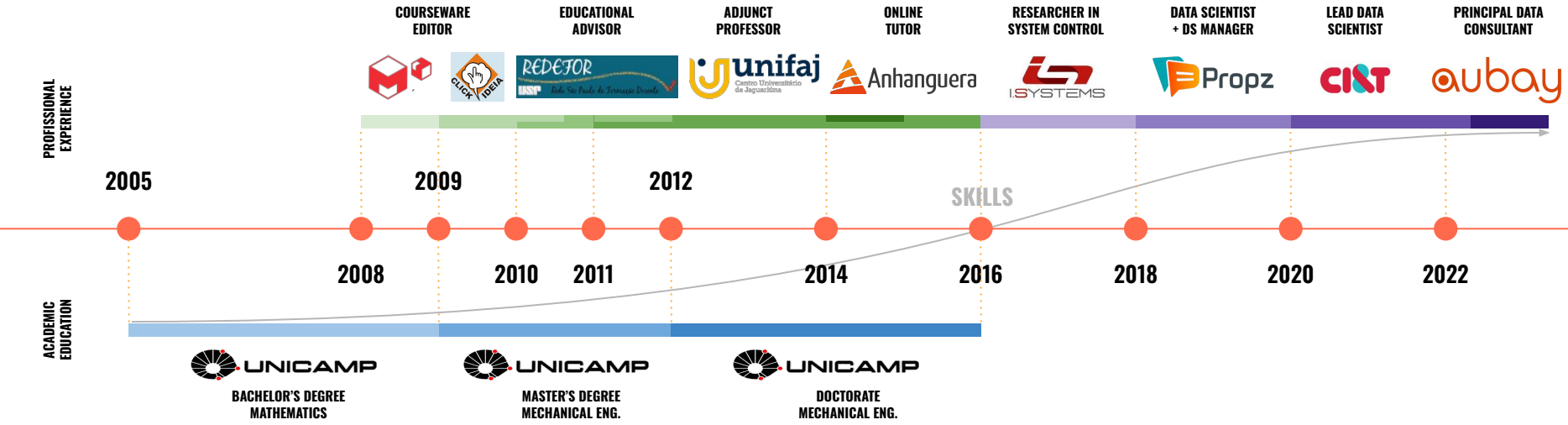
# Customer Complaints Email Classification

## Case Study | ML Engineer

Luis Rodrigues

luis-rodrigues-phd

# PROFESSIONAL TRAJECTORY OF LUIS RODRIGUES




PROJECT EVOLUTION

**Propz**

**MARKETING/CRM PROJECT**
EDA, Customer Segmentation, KPI

**AI PROJECT**
Big Data, Recommender Systems

**CI&T**

**INNOVATION PROJECT**
A/B Testing, Data Analysis

**INNOVATION PROJECT**
NLP, A/B Testing, OKR

**AI PROJECT**
Recommender Systems

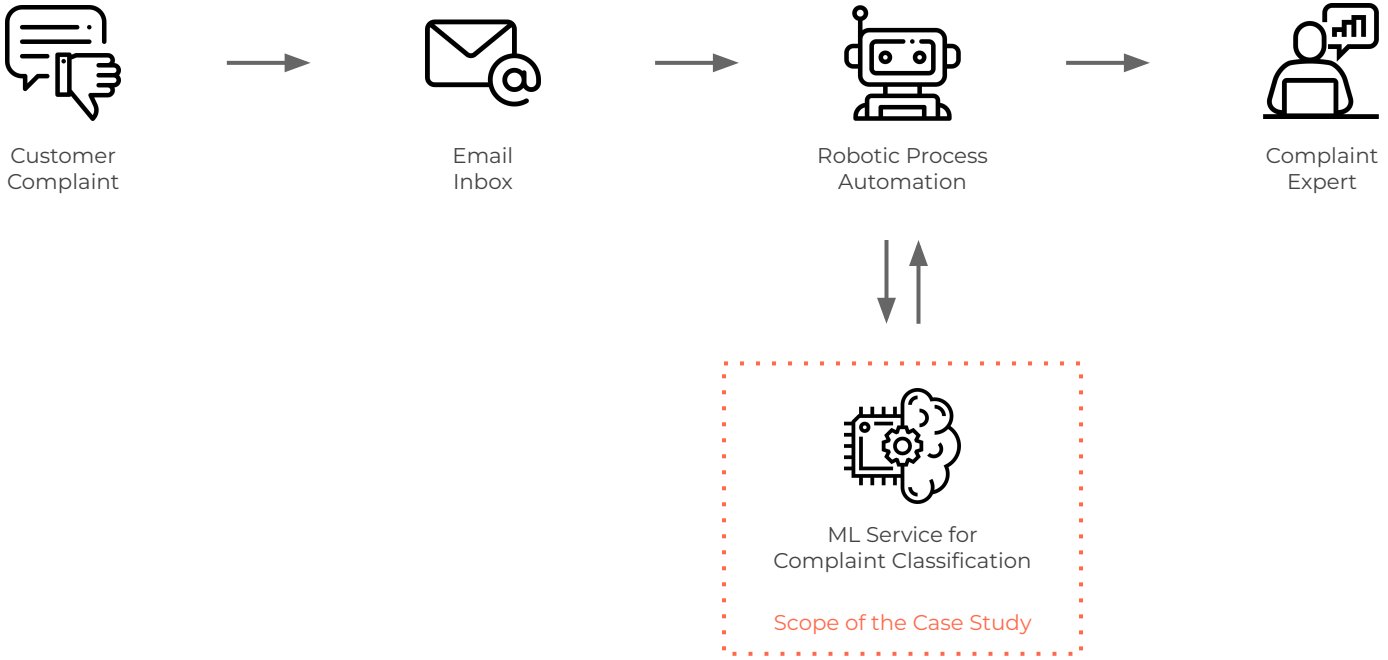
**aubay**

**AI PROJECT**
NLP, LLM, (Near) Real-Time Predictions

# 01

# CASE STUDY

# CASE STUDY | CUSTOMER COMPLAINTS EMAIL CLASSIFICATION




Implement an ML inference pipeline and deploy a REST endpoint that consumes the trained ML model

# CASE STUDY | CUSTOMER COMPLAINTS EMAIL CLASSIFICATION

## About Dataset

### Context

The Consumer Financial Protection Bureau (CFPB) is a federal U.S. agency that acts as a mediator when disputes arise between financial institutions and consumers. Via a web form, consumers can send the agency a narrative of their dispute. An NLP model would make the classification of complaints and their routing to the appropriate teams more efficient than manually tagged complaints.

### Content

A data file was downloaded directly from the CFPB website for training and testing the model. It included one year's worth of data (March 2020 to March 2021). Later in the project, I used an API to download up-to-the-minute data to verify the model's performance.

Each submission was tagged with one of nine financial product classes. Because of similarities between certain classes as well some class imbalances, I consolidated them into five classes: credit reporting, debt collection, mortgages and loans (includes car loans, payday loans, student loans, etc.), credit cards, retail banking (includes checking/savings accounts, as well as money transfers, Venmo, etc.).

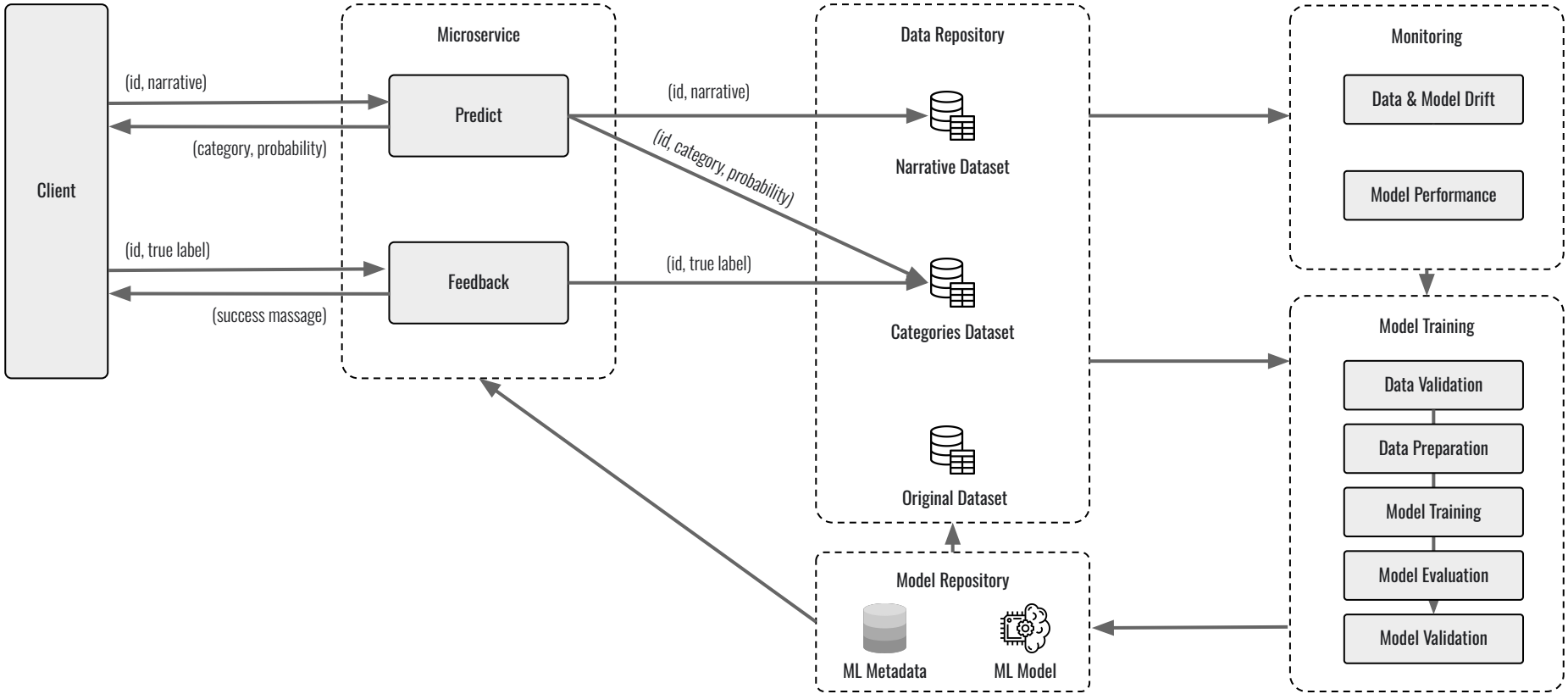
After data cleaning, the dataset consisted of around 162,400 consumer submissions containing narratives. The dataset was still imbalanced, with 56% in the credit reporting class, and the remainder roughly equally distributed (between 8% and 14%) among the remaining classes.

https://www.kaggle.com/datasets/shashwatwork/consume-complaints-dataset-fo-nlp

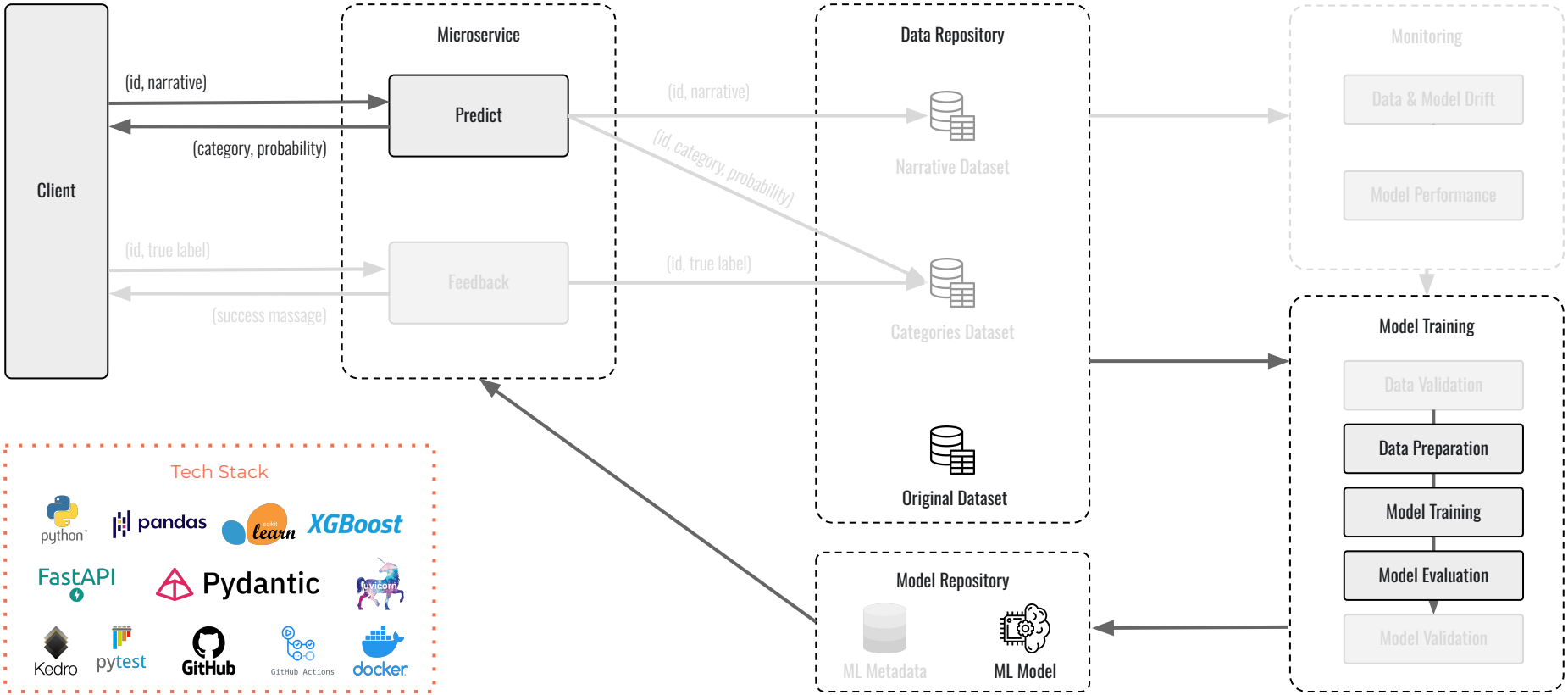
# 02

# SOLUTION

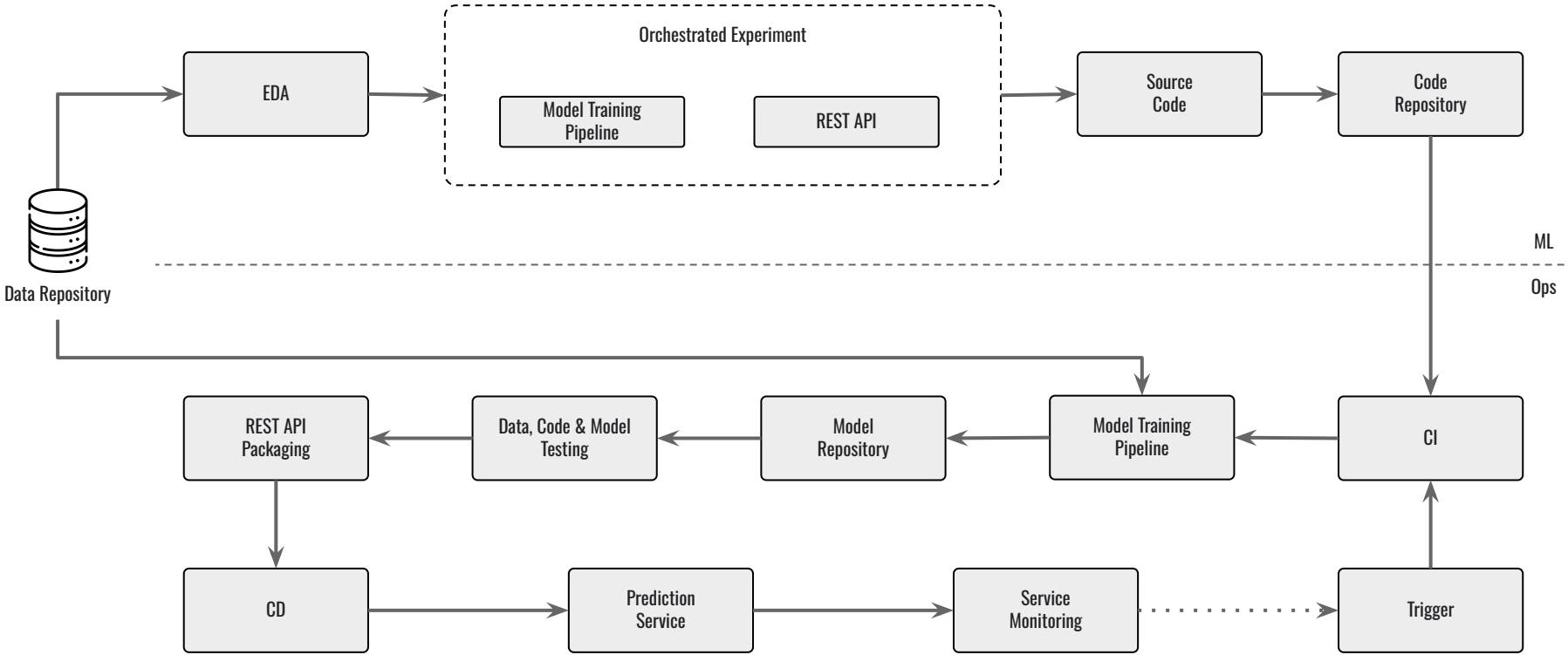
# SOLUTION ARCHITECTURE | PROPOSAL



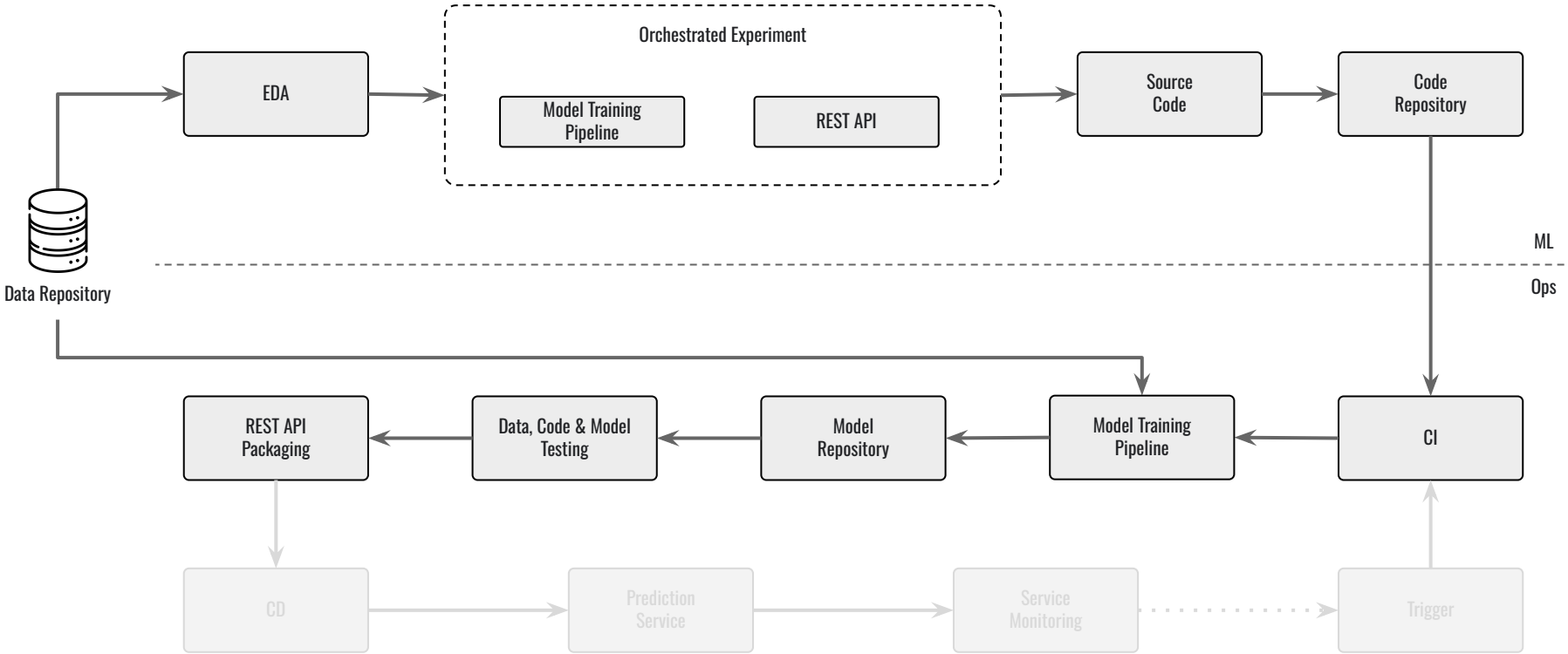
# SOLUTION ARCHITECTURE | DEVELOPED



# CI/CD/CT PIPELINE | PROPOSAL



# CI/CD/CT PIPELINE | DEVELOPED



# DEMONSTRATION

1. ML model prototype in notebook
2. Kedro architecture overview
3. Development setup
4. Productionized model training pipeline
5. Train a model and show logs and artifacts
6. Productionized model inference pipeline
7. Make a prediction and show logs
8. REST API to serve online predictions
9. Show the prediction service running locally
10. Show and run data, code and model tests
11. GitHub Actions workflow to check PRs
12. Create a PR to show the workflow running
13. GitHub Actions workflow for deployment
14. Merge the PR to show the workflow running
15. Dockerfile to package the prediction service
16. Train the model and serve prediction using the container

# THANK YOU