Napoleon IT

Как небольшой компании запустить MLOps?

Наша повестка

N IT

01
Введение

О компании
Что было
и почему пошли в
MLOps

02
Итерация 0

Ресерч
Стало
Вывод

03
Итерация I

Ресерч
Стало
Вывод

04
Итерация II

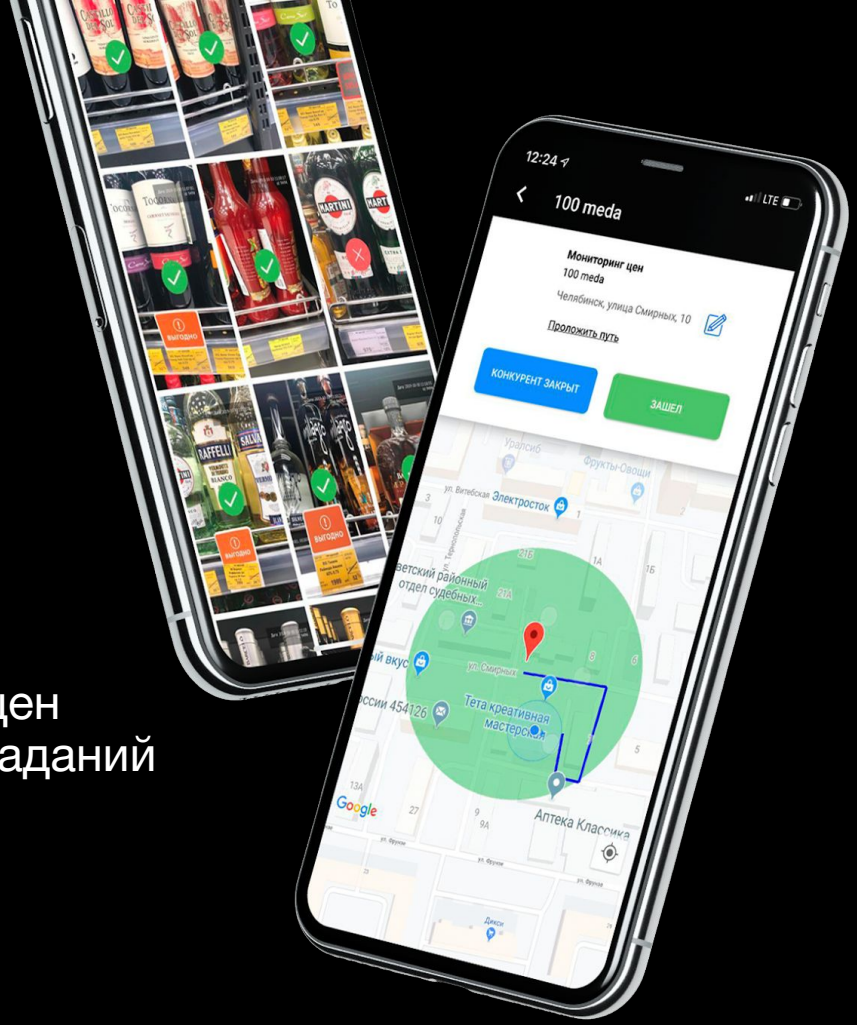
Ресерч
Стало
Вывод

05
Заключение

Что делать
или roadmap
по кварталам

CASE

Продуктовый ритейлер



Проблема

Трудоемкость сбора и обработки информации о ценах конкурентов в магазинах

Решение

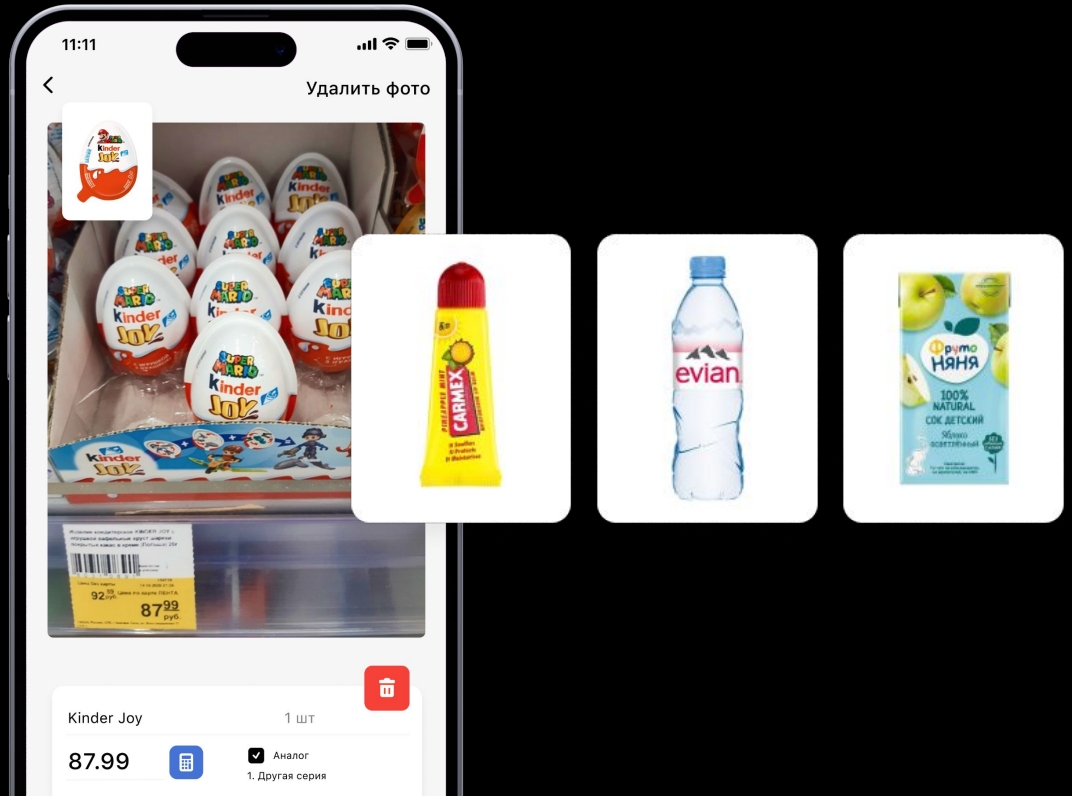
Приложение для оффлайн-мониторинга цен конкурентов с функцией планировщика заданий

Супервайзер ставит работнику задачу по мониторингу цен на отдельные товарные позиции в магазинах конкурентов

Пользователь делает фотографии товара с ценником или всего стеллажа. Они автоматически распознаются, заносятся в реестр и передаются супервайзеру в режиме реального времени

CASE

Показатели бизнес-кейса



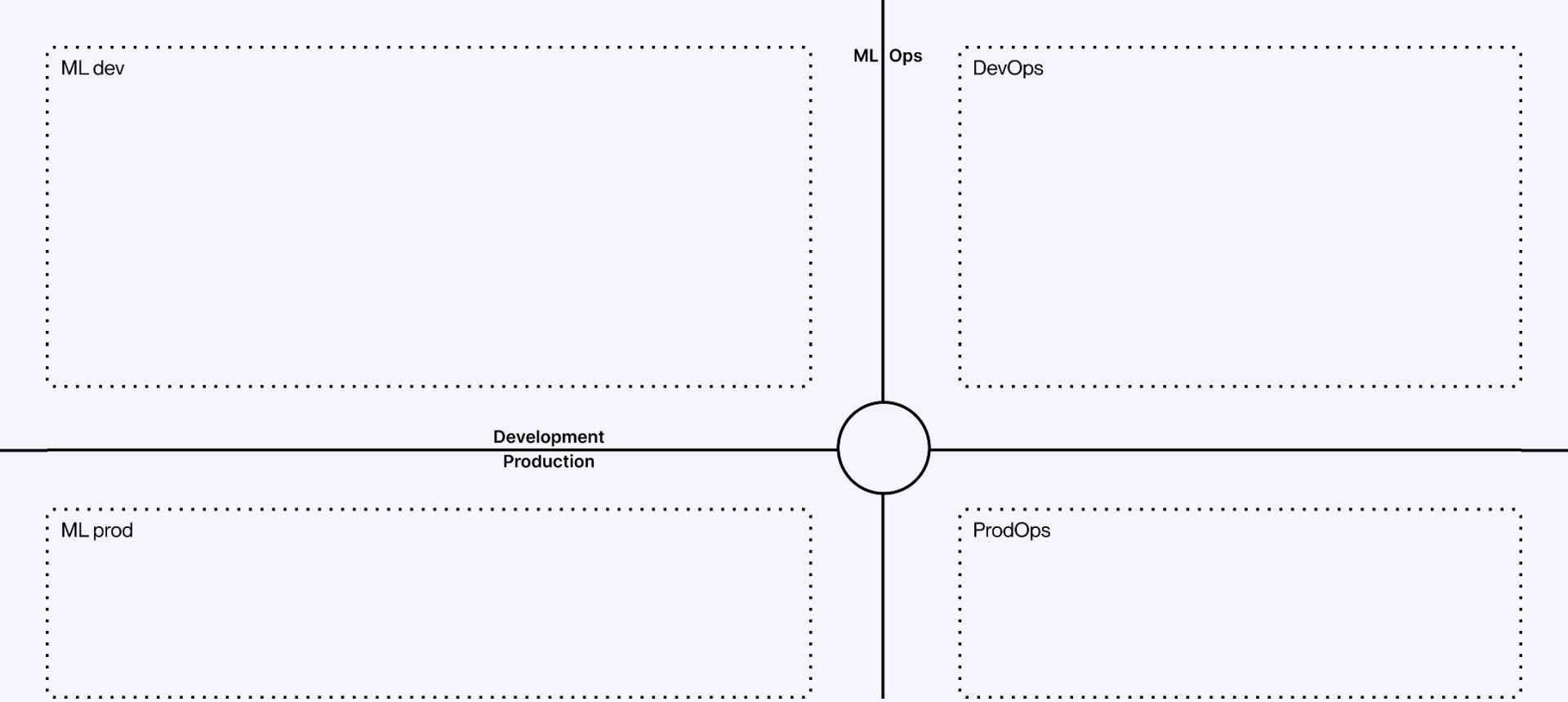
87 → 33 мин

сокращение времени мониторинга в расчете на 100 SKU

20%

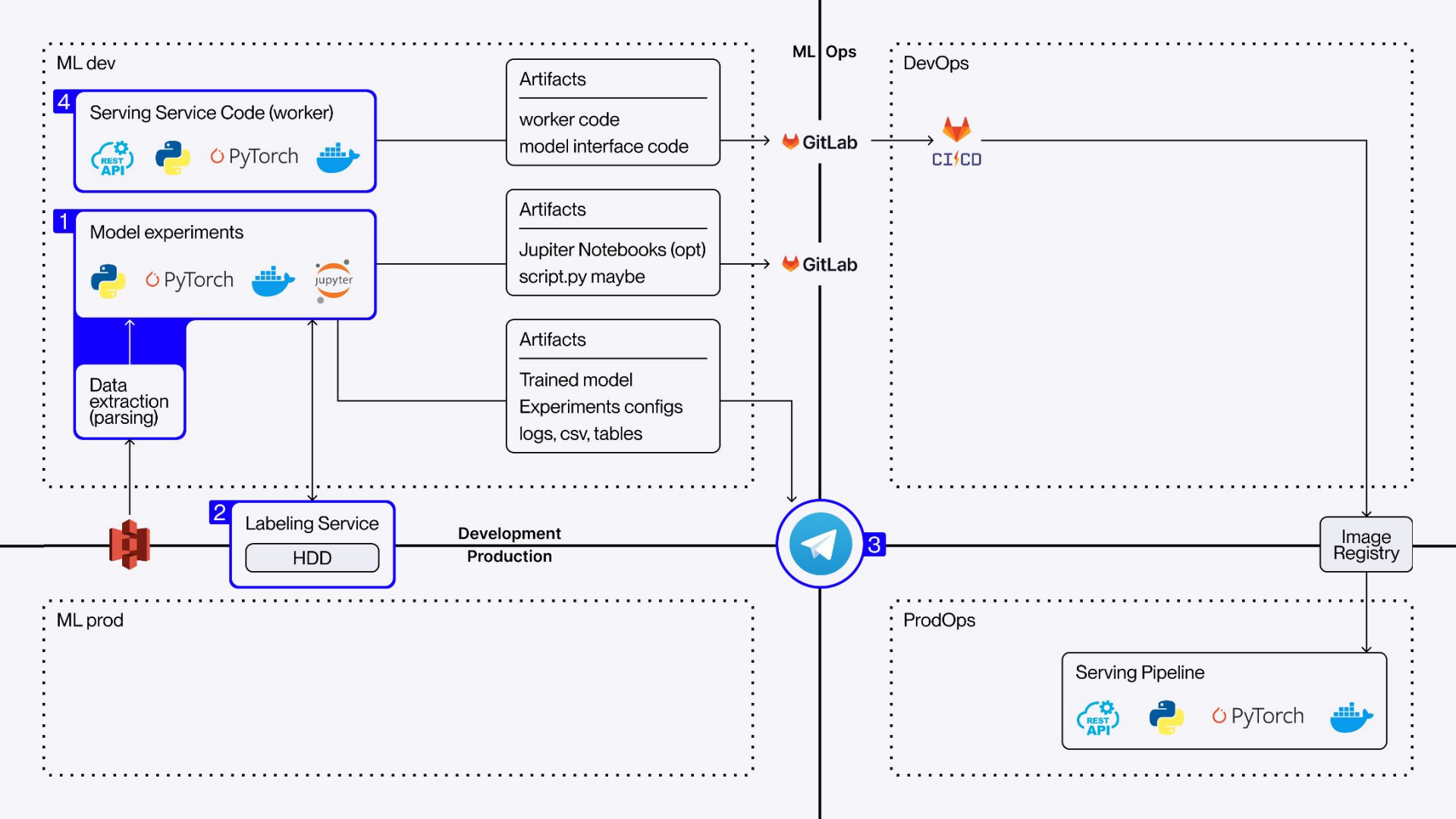
рост точности собираемых данных по ценам конкурентов

увеличение охвата рынка конкурентных магазинов в ходе мониторинга цен



MLOps в билайн: как катить машинное обучение в production без ML-инженеров

Построение MLOps платформы. Как мы обуздали хаос в головах и технике


ML dev
4
Serving Service Code (worker)
REST API
PyTorch
1
Model experiments
PyTorch
jupyter
Data extraction (parsing)
Artifacts
worker code
model interface code
Artifacts
Jupiter Notebooks (opt)
script.py maybe
Artifacts
Trained model
Experiments configs
logs, csv, tables
ML Ops
GitLab
GitLab
DevOps
CI/CD
2
Labeling Service
HDD
Development
Production
3
Image Registry
ML prod
ProdOps
Serving Pipeline
REST API
PyTorch

N IT

Почему пошли в MLOps

→ Увольнение разработчика

→ А где та модель, которую мы получили 2 месяца назад?

→ Ой, Я случайно сделал sudo rm-rf на датасет

→ Проблема с утилизацией железа и деплоем моделей

90%
кода лежит локально

90%
артефактов лежат локально

100%
ручной однотипной работы для релиза

0%?
проекта удовлетворяет концепциям mlops

N IT

Интеграция 0

Ресерч

DevOps
MLOps
CloudOps
AIOps
DataOps
ITOps
RevOps
FinOps
HROps
LegalOps
BizOps
LLMOps
ChatOps
NoOps

MLOps — методология оптимизации, ускорения и автоматизации процесса разработки и доставки решений Машинного обучения

Ресерч. Требования

N IT

01

Нужно что-то, чтобы хранить и версионировать артефакты

02

Нужно что-то, что сможет выступать как оркестратор для пайплайнов

03

Нужно что-то, чтобы хранить и версионировать данные

04

Нужно что-то, что предоставляет функционал разметки данных

Ресерч. Результат

N IT

mlflow™

Логирование экспериментов и моделей

DVC

Версионирование данных

Apache Airflow

Оркестратор

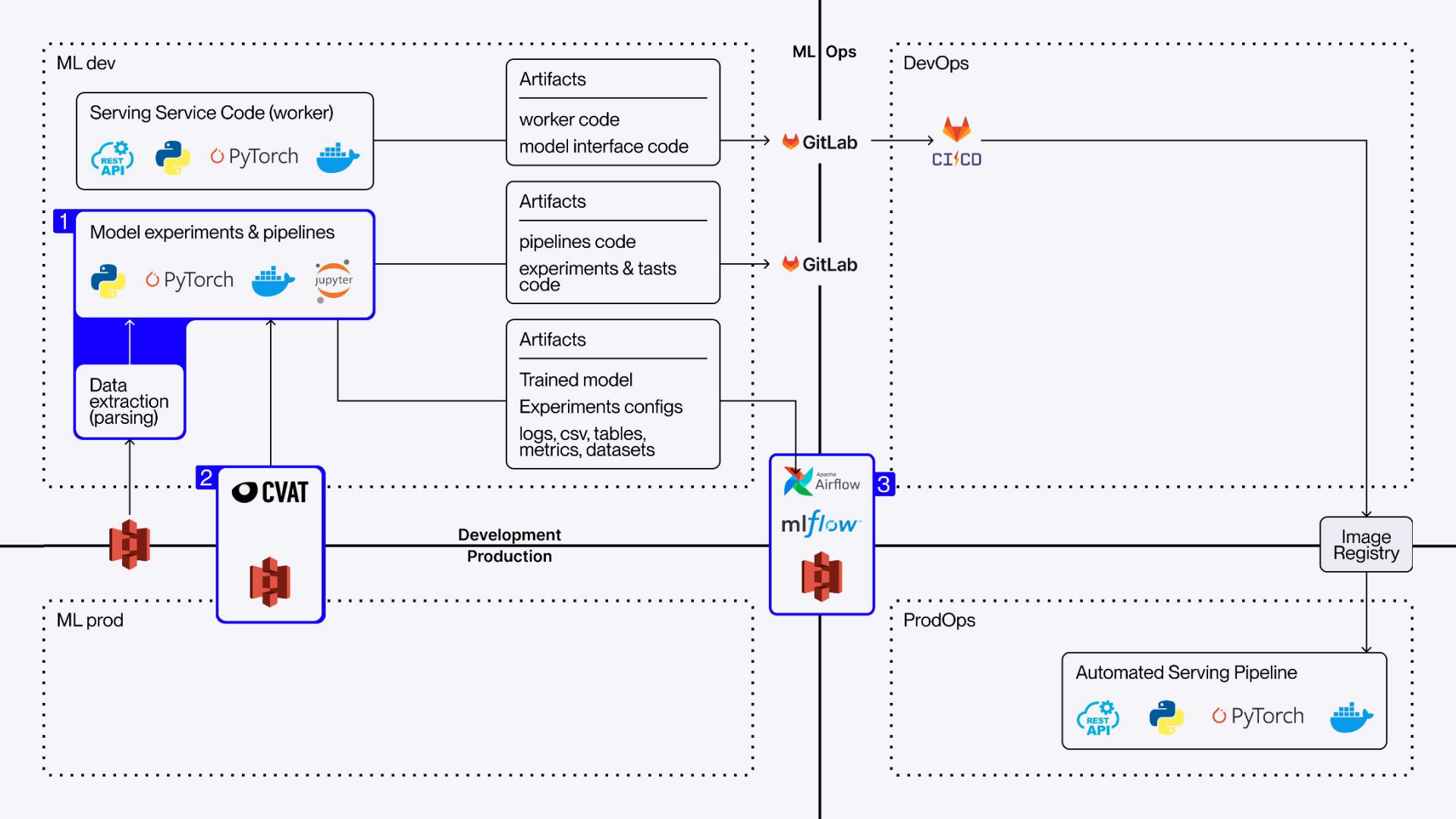
CVAT

Сервис разметки

MLOps: Continuous delivery and automation pipelines in machine learning

The 2024 Mad (Machine Learning, Artificial Intelligence & Data) Landscape

Стало. Инструменты


ML dev
Serving Service Code (worker)
REST API
PyTorch
Artifacts
worker code
model interface code
1
Model experiments & pipelines
PyTorch
jupyter
Artifacts
pipelines code
experiments & tasts code
Data extraction (parsing)
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
2
CVAT
ML Ops
GitLab
GitLab
Apache Airflow
mlflow
3
DevOps
CI/CD
Development
Production
Image Registry
ML prod
ProdOps
Automated Serving Pipeline
REST API
PyTorch

Вывод

→ Единое окно: трудности в эксплуатации и поддержке такой инфраструктуры

→ Сложности в масштабировании — нужно писать кучу обвязок вокруг этих инструментов (возвращаемся к проблеме выше)

→ Инструментарий бигтеха не всегда лучший выбор для компаний/команд маленького/среднего размера

90% → 0%
кода лежит локально

90% → 0%
артефактов лежат локально

100% → 60%
ручной однотипной работы для релиза

0% → 100%?
проекта удовлетворяет концепциям MLOps

N IT

Ресерч

N IT

01 Изменение метрик, по которым отслеживаем результат внедрения MLOps

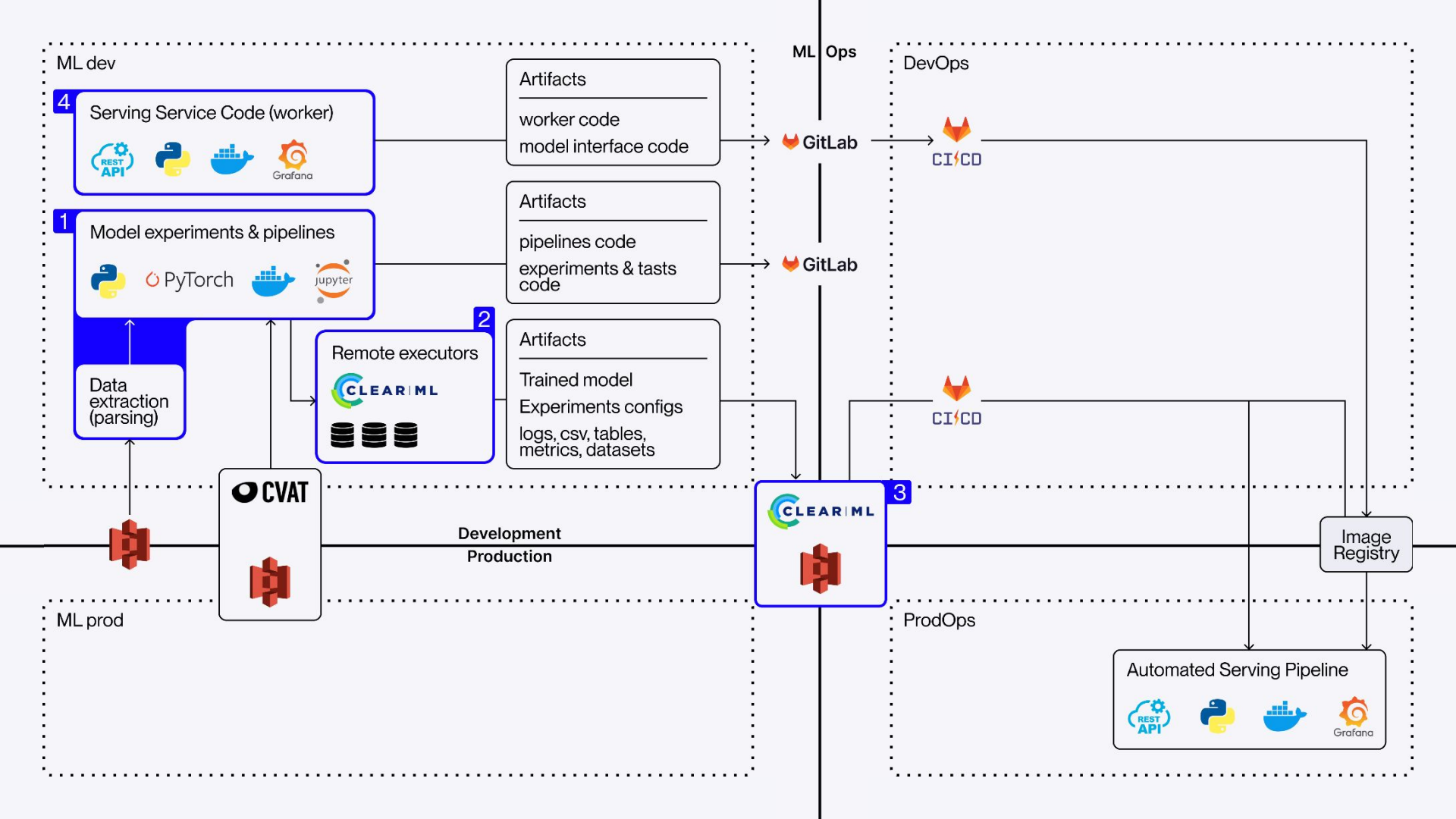
02 Более точечный ресерч с учетом опыта

→ Более ML-ый оркестратор

→ Единое окно для рабочего места ML специалиста

Ресерч

	ML Flow	Clear ML	Wandb	DVC studio	V7
Интерфейс	-	+	+	+ -	+
On Premise	+	+	+	+	-
Enterprise version	-	+	+	+	+ только платная
RBAC	-	- только платная	-	- +	+
Пайплайны	+ -	+	-	-	+
Remote execution	-	+	-	-	-
Мониторинг	-	+	-	-	+
Версионирование	-	+	+	+	+
Разметка	-	+ enterprise	-	-	+


ML dev
4
Serving Service Code (worker)
REST API
Grafana
1
Model experiments & pipelines
PyTorch
jupyter
Data extraction (parsing)
2
Remote executors
CLEARML
Artifacts
worker code
model interface code
Artifacts
pipelines code
experiments & tasts code
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
ML Ops
GitLab
GitLab
DevOps
CI/CD
CI/CD
CVAT
Development
Production
3
CLEARML
Image Registry
ML prod
ProdOps
Automated Serving Pipeline
REST API
Grafana

Вывод

→ ClearML прост в использовании и имеет большой функционал из коробки

→ Обернуть в скрипты + remote executors + release_pipeline уменьшаем время релиза и увеличиваем утилизацию железа на 35%

→ CI/CD уменьшил кол-во факапов деплоя моделей на 30%

0% → 25%
уменьшено время релиза

–
уменьшено время демо

0% → 35%
соответствие уровню MLOps lvl 2

N IT

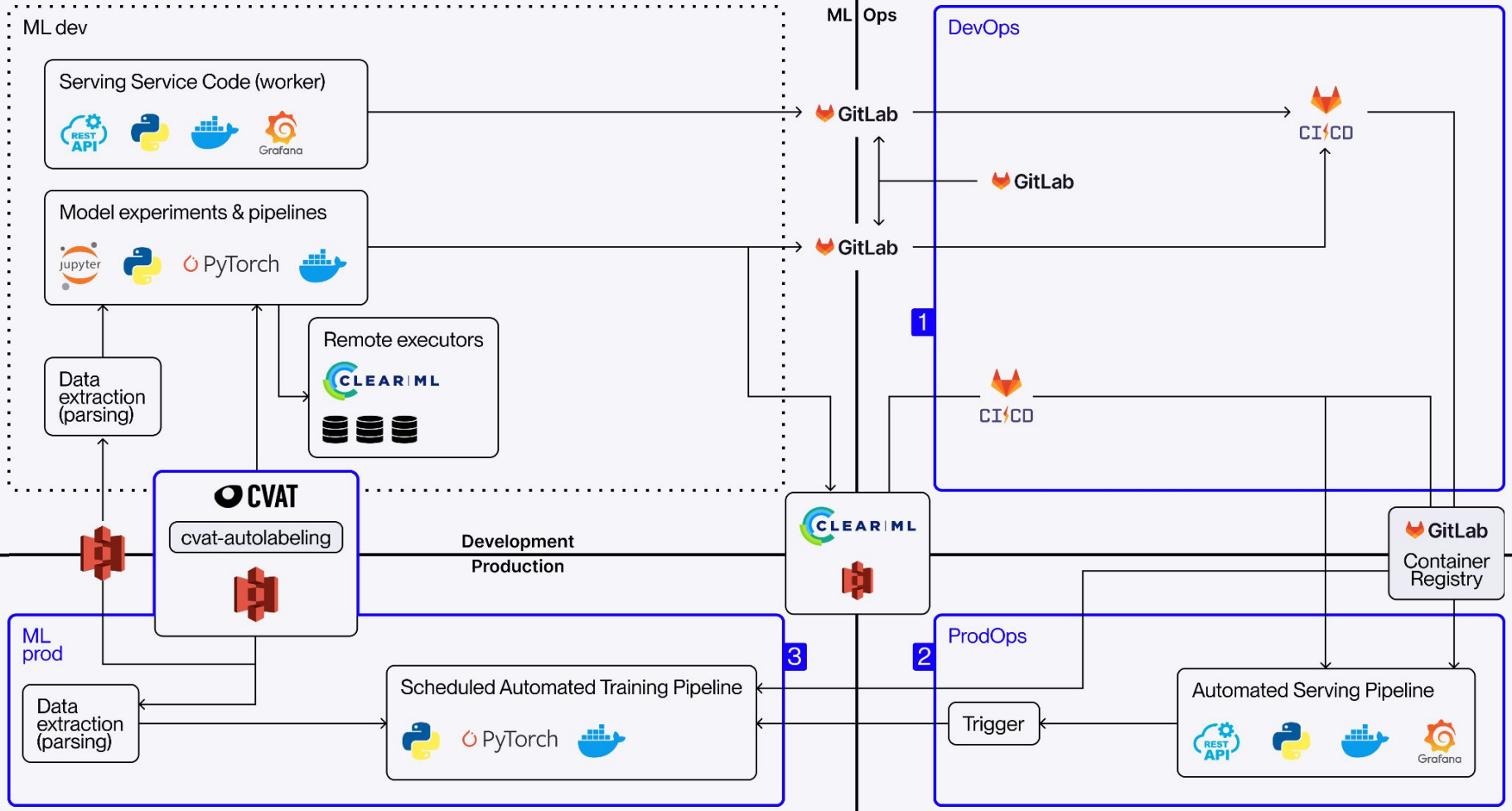
Ресерч

N IT

→ А как организовать рабочее место не только ML Engineer но и Data Analyst?

→ С dev разобрались, а что делать с prod?

→ Можем ли мы еще ускорить процесс обучения?


ML dev
Serving Service Code (worker)
REST API
Grafana
Model experiments & pipelines
jupyter
PyTorch
Remote executors
CLEARML
Data extraction (parsing)
CVAT
cvat-autolabeling
ML Ops
GitLab
GitLab
GitLab
DevOps
CI/CD
CI/CD
1
Development
Production
CLEARML
GitLab
Container Registry
ML prod
Data extraction (parsing)
Scheduled Automated Training Pipeline
PyTorch
3
2
ProdOps
Trigger
Automated Serving Pipeline
REST API
Grafana

Выводы

→ Авторазметка CVAT-а экономит ресурсы и уменьшает время разметки => релиза

→ Добавление шедулеров на пайплайны уменьшает время релизов и загруженность разработчиков

→ Даже костыльные триггеры на метрики в связке с автоматизацией дают прирост по метрикам

25% → 65%
уменьшено время релиза

0% → 20%
уменьшено время демо

20% → 65%
соответствие уровню MLOps lvl 2

Планы

→ Поддержка и развитие текущих инструментов

→ Ресерч, ресерч и еще раз ресерч

→ Оптимизация процесса деплоя и сервисов-оберток моделей (Seldon Core?)

→ Рабочее место дата аналитика и дашборды (SuperSet? JupiterHub?)

→ Доработка мониторингов и триггеров (Evidently?)

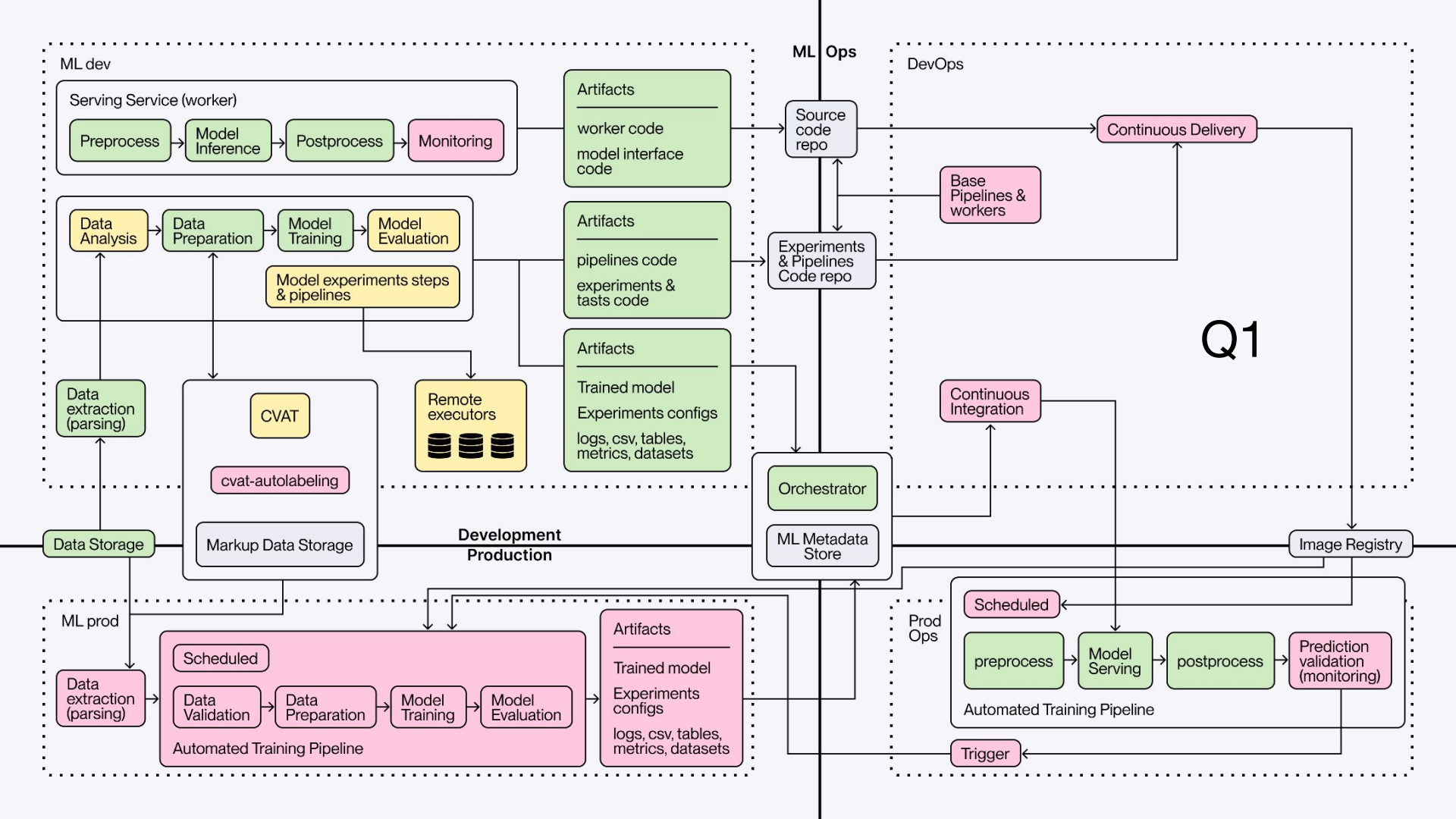
→ Доработка валидации данных (Great Expectations?)

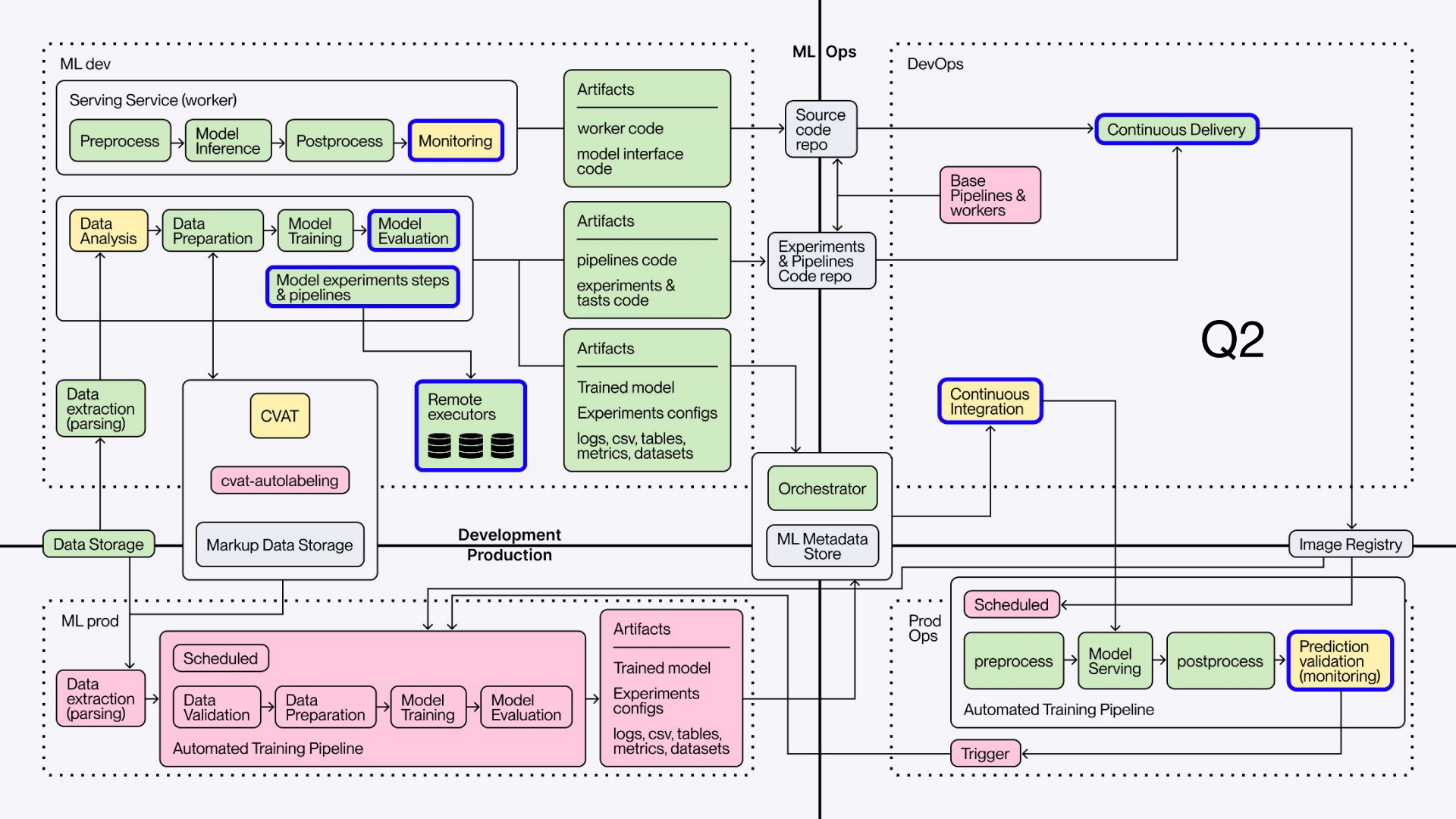
N IT

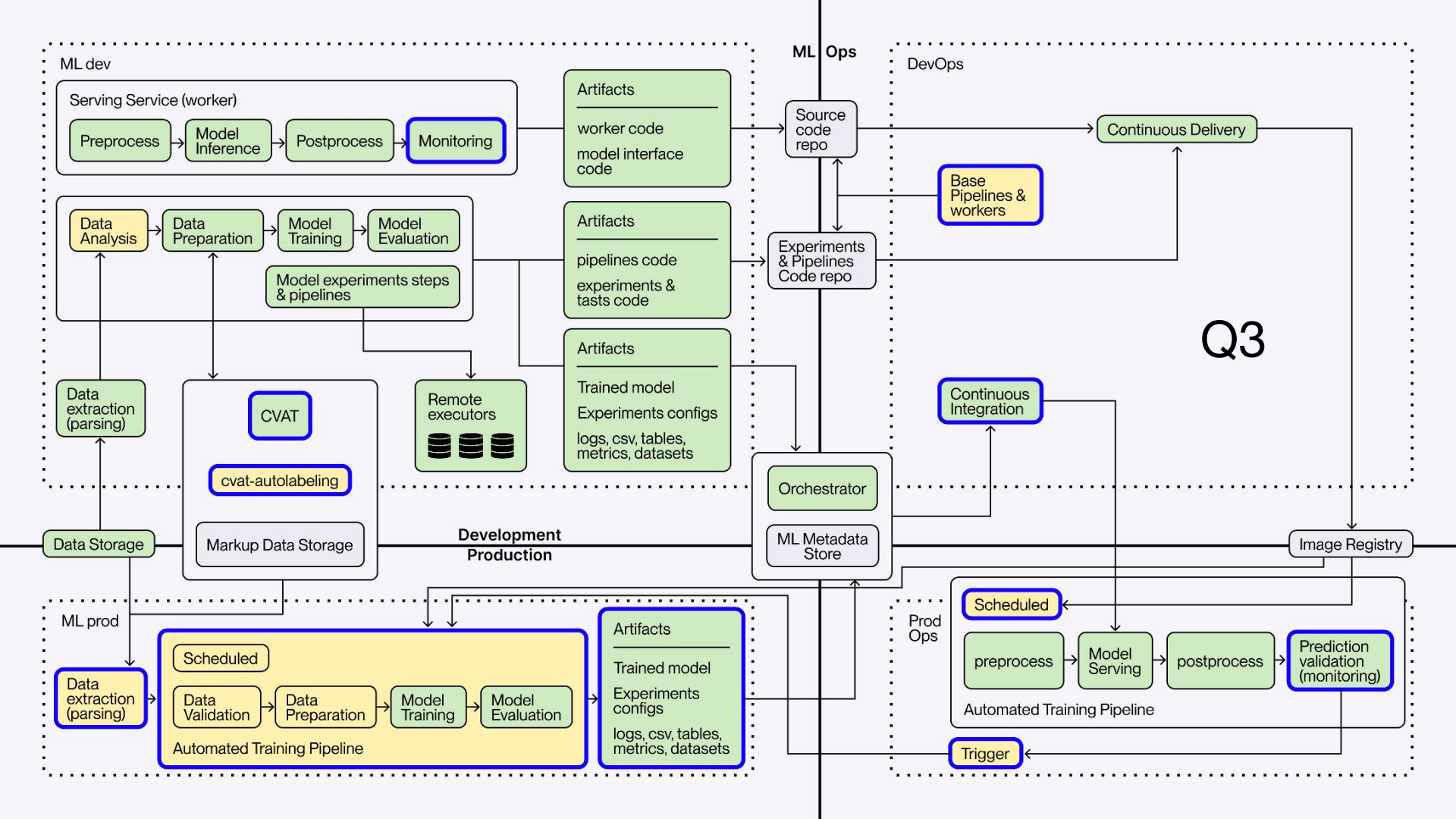
Что делать или RoadMap по кварталам

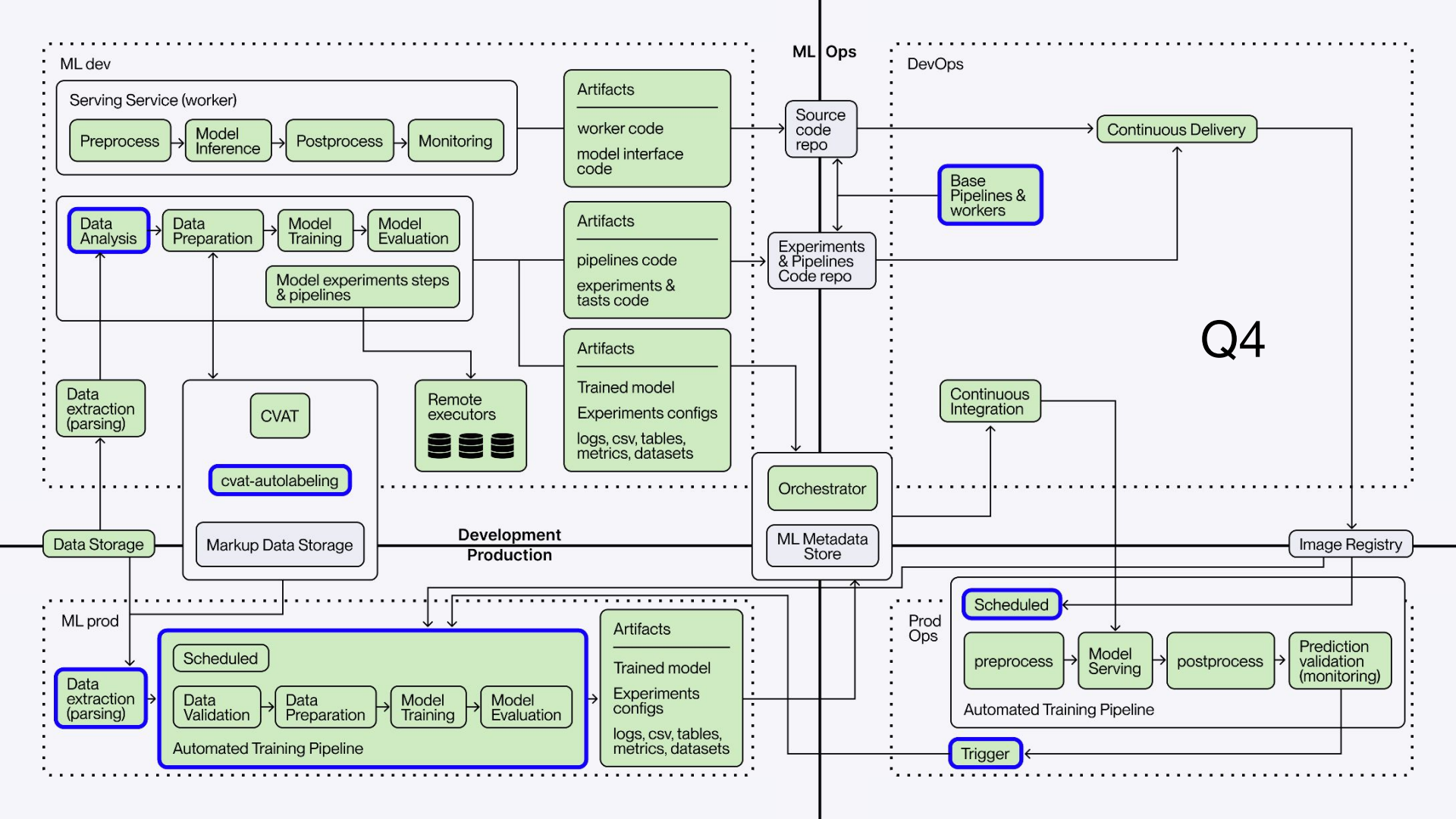
N IT

→ Проекты с неструктурированными данными
→ 2 ML Engineer-a
→ 1 DevOps
→ Team Lead
→ Бешенное желание страдать меньше


ML dev
Serving Service (worker)
Preprocess
Model Inference
Postprocess
Monitoring
Artifacts
worker code
model interface code
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasts code
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Data extraction (parsing)
CVAT
cvat-autolabeling
Markup Data Storage
Remote executors
Data Storage
ML Ops
Source code repo
Experiments & Pipelines Code repo
Orchestrator
ML Metadata Store
DevOps
Continuous Delivery
Base Pipelines & workers
Continuous Integration
Q1
Image Registry
Development
Production
ML prod
Data extraction (parsing)
Scheduled
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Prod Ops
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Training Pipeline
Trigger


ML dev
Serving Service (worker)
Preprocess
Model Inference
Postprocess
Monitoring
Artifacts
worker code
model interface code
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasts code
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Data extraction (parsing)
CVAT
cvat-autolabeling
Markup Data Storage
Remote executors
Data Storage
ML Ops
Source code repo
Experiments & Pipelines Code repo
Orchestrator
ML Metadata Store
DevOps
Continuous Delivery
Base Pipelines & workers
Continuous Integration
Q2
Image Registry
Development
Production
ML prod
Data extraction (parsing)
Scheduled
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Prod Ops
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Training Pipeline
Trigger


ML dev
Serving Service (worker)
Preprocess
Model Inference
Postprocess
Monitoring
Artifacts
worker code
model interface code
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasts code
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Data extraction (parsing)
CVAT
cvat-autolabeling
Markup Data Storage
Remote executors
Data Storage
ML Ops
Source code repo
Experiments & Pipelines Code repo
Base Pipelines & workers
Orchestrator
ML Metadata Store
DevOps
Continuous Delivery
Continuous Integration
Q3
Development
Production
Image Registry
ML prod
Scheduled
Data extraction (parsing)
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Prod Ops
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Training Pipeline
Trigger


ML dev
Serving Service (worker)
Preprocess
Model Inference
Postprocess
Monitoring
Artifacts
worker code
model interface code
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasts code
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Data extraction (parsing)
CVAT
cvat-autolabeling
Markup Data Storage
Remote executors
Data Storage
ML Ops
Source code repo
Experiments & Pipelines Code repo
Orchestrator
ML Metadata Store
DevOps
Continuous Delivery
Base Pipelines & workers
Q4
Continuous Integration
Image Registry
Development
Production
ML prod
Data extraction (parsing)
Scheduled
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Prod Ops
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Training Pipeline
Trigger

N IT

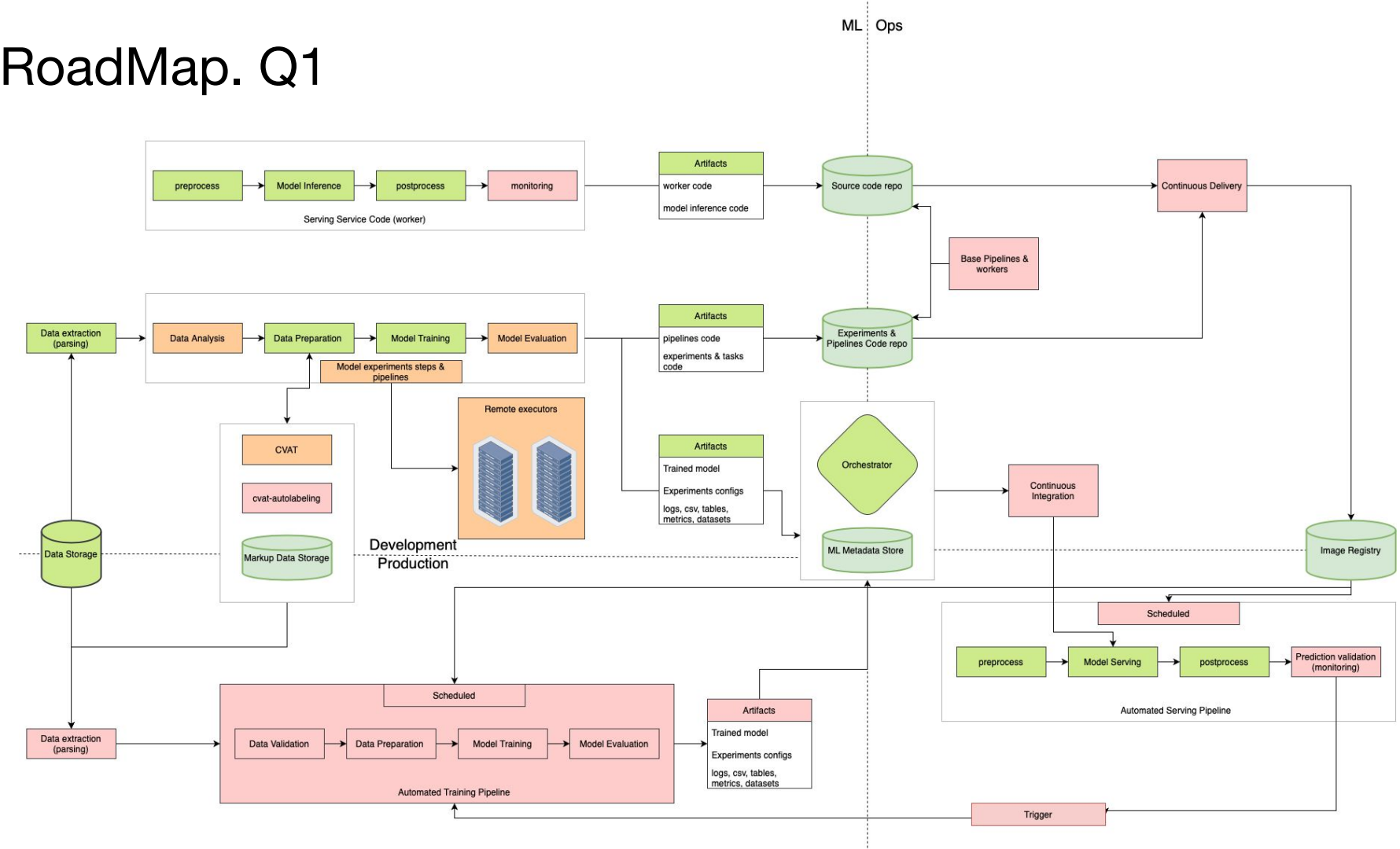
Игорь Терёхин

Руководитель отдела
ML Napoleon IT

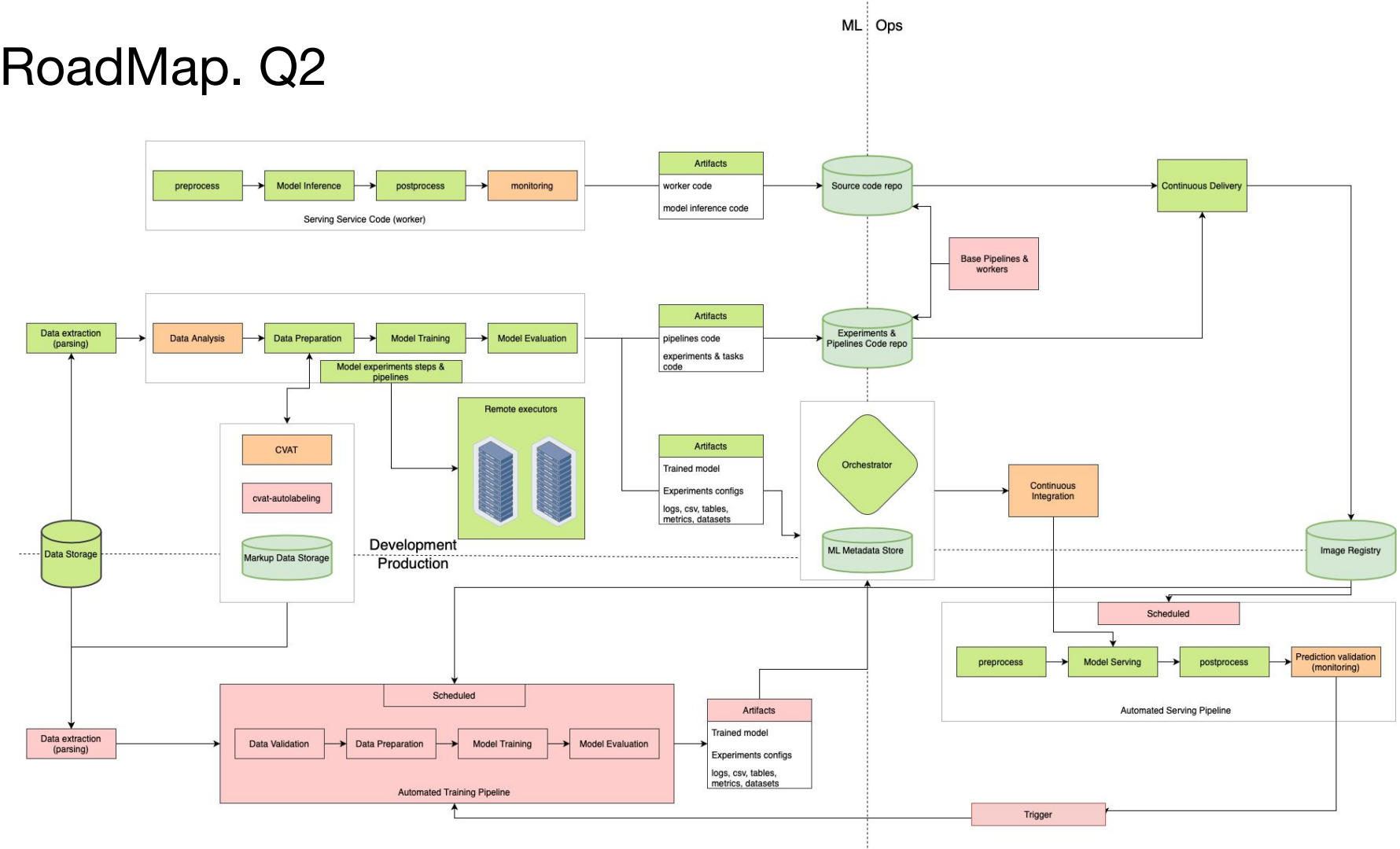
@spannenberger

i.terekhin@napoleonit.ru

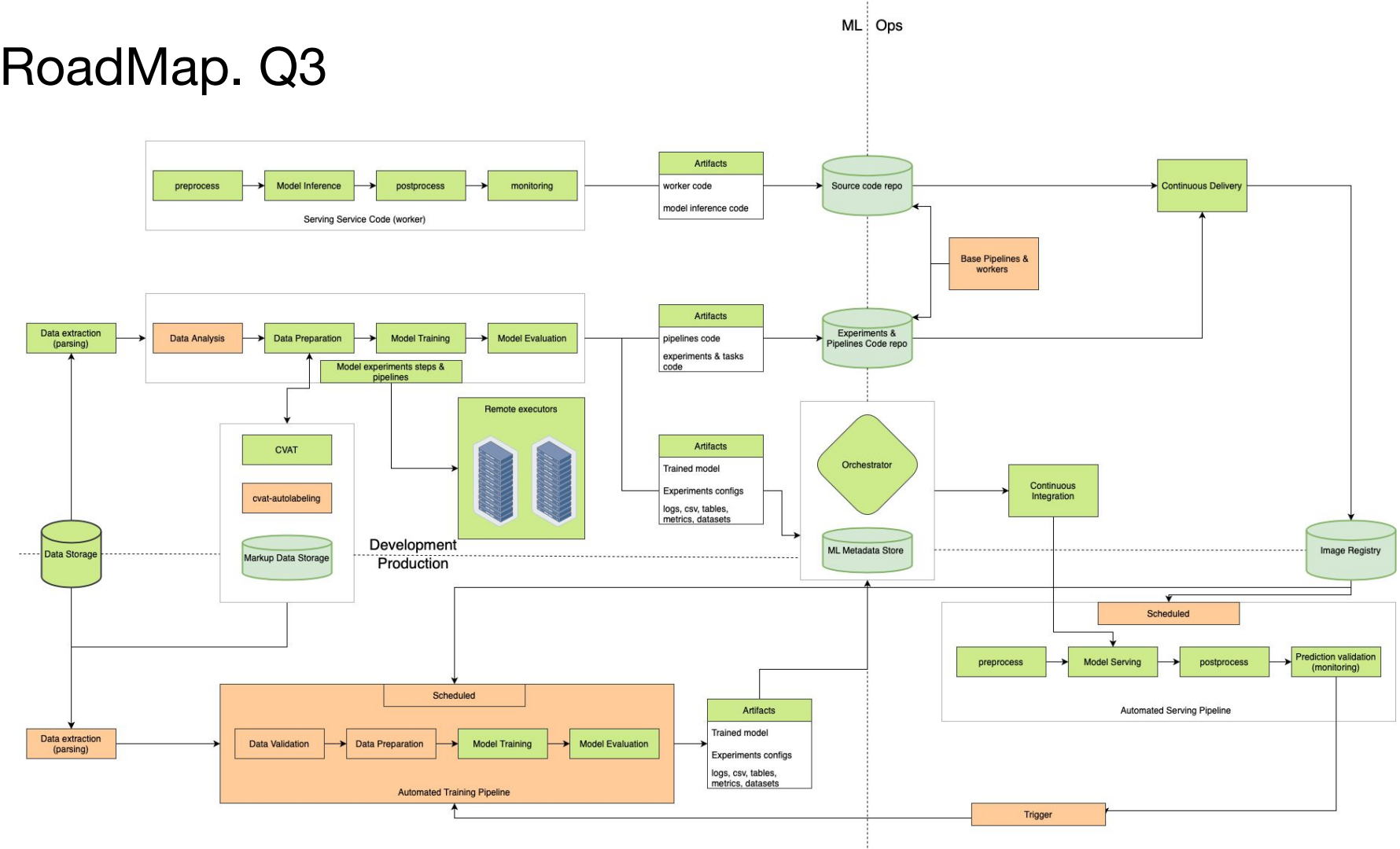
RoadMap. Q1


ML Ops
preprocess
Model Inference
postprocess
monitoring
Serving Service Code (worker)
Artifacts
worker code
model inference code
Source code repo
Continuous Delivery
Base Pipelines & workers
Data extraction (parsing)
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasks code
Experiments & Pipelines Code repo
Remote executors
CVAT
cvat-autolabeling
Markup Data Storage
Data Storage
Development
Production
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Orchestrator
ML Metadata Store
Continuous Integration
Image Registry
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Serving Pipeline
Data extraction (parsing)
Scheduled
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Trigger

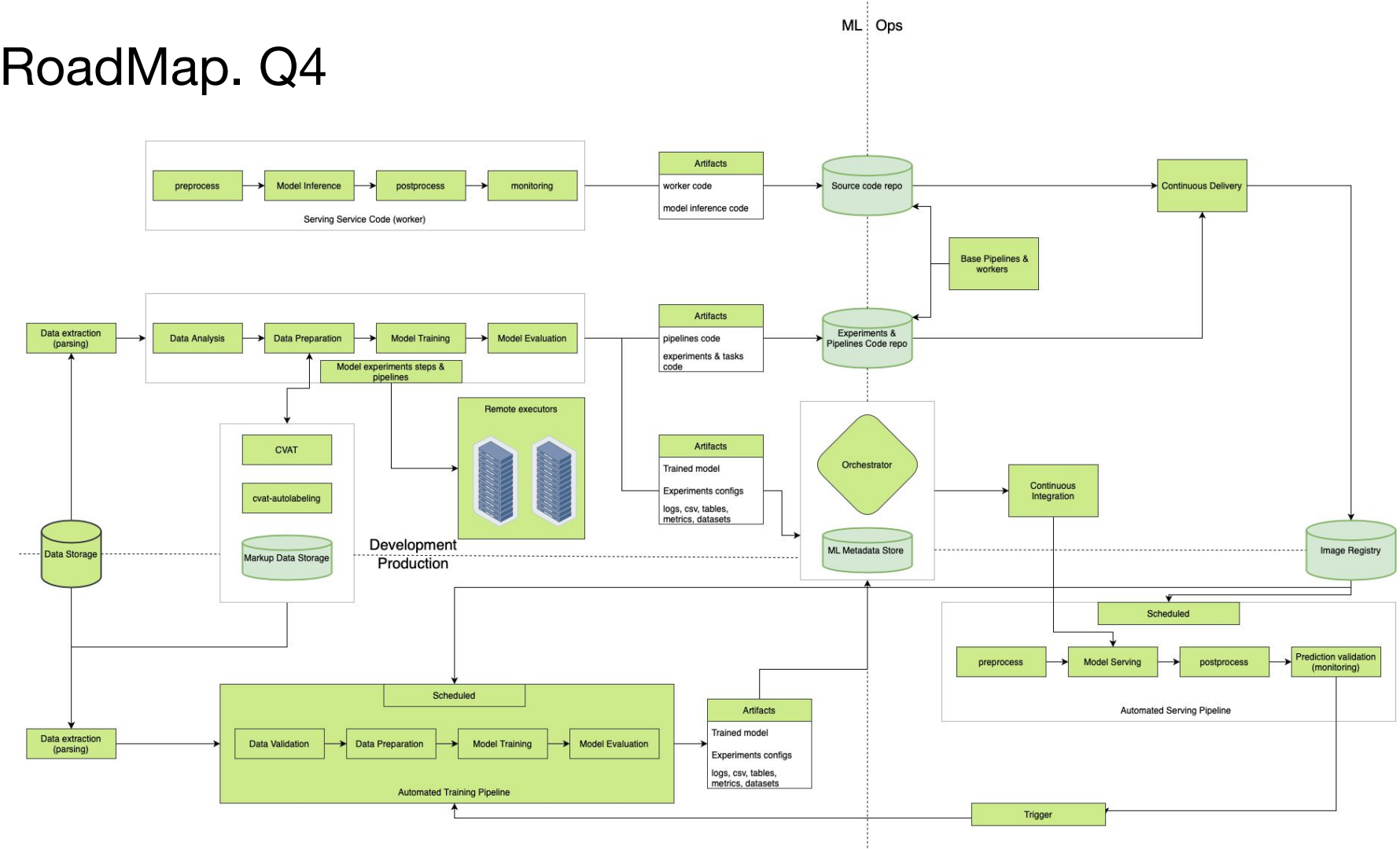
RoadMap. Q2


ML Ops
preprocess
Model Inference
postprocess
monitoring
Serving Service Code (worker)
Artifacts
worker code
model inference code
Source code repo
Continuous Delivery
Base Pipelines & workers
Data extraction (parsing)
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasks code
Experiments & Pipelines Code repo
Remote executors
CVAT
cvat-autolabeling
Markup Data Storage
Data Storage
Development
Production
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Orchestrator
ML Metadata Store
Continuous Integration
Image Registry
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Serving Pipeline
Data extraction (parsing)
Scheduled
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Trigger

RoadMap. Q3


ML Ops
preprocess
Model Inference
postprocess
monitoring
Serving Service Code (worker)
Artifacts
worker code
model inference code
Source code repo
Continuous Delivery
Base Pipelines & workers
Data extraction (parsing)
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasks code
Experiments & Pipelines Code repo
Remote executors
CVAT
cvat-autolabeling
Markup Data Storage
Data Storage
Development
Production
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Orchestrator
ML Metadata Store
Continuous Integration
Image Registry
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Serving Pipeline
Data extraction (parsing)
Scheduled
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Trigger

RoadMap. Q4


ML Ops
preprocess
Model Inference
postprocess
monitoring
Serving Service Code (worker)
Artifacts
worker code
model inference code
Source code repo
Continuous Delivery
Base Pipelines & workers
Data extraction (parsing)
Data Analysis
Data Preparation
Model Training
Model Evaluation
Model experiments steps & pipelines
Artifacts
pipelines code
experiments & tasks code
Experiments & Pipelines Code repo
Remote executors
CVAT
cvat-autolabeling
Markup Data Storage
Data Storage
Development
Production
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Orchestrator
ML Metadata Store
Continuous Integration
Image Registry
Scheduled
preprocess
Model Serving
postprocess
Prediction validation (monitoring)
Automated Serving Pipeline
Data extraction (parsing)
Scheduled
Data Validation
Data Preparation
Model Training
Model Evaluation
Automated Training Pipeline
Artifacts
Trained model
Experiments configs
logs, csv, tables, metrics, datasets
Trigger