Building a **Real-Time** ML System Together

Session 1

2024-09-16

1. Problem framing → From a business problem to an ML system

1. Problem framing → From a business problem to an ML system

2. ML System design

- 1. Problem framing → From a business problem to an ML system
- 2. ML System design
 - **♦** Feature-Training-Inference Pipelines

- 1. Problem framing → From a business problem to an ML system
- 2. ML System design
 - ◆ Feature-Training-Inference Pipelines
 - Infrastructure

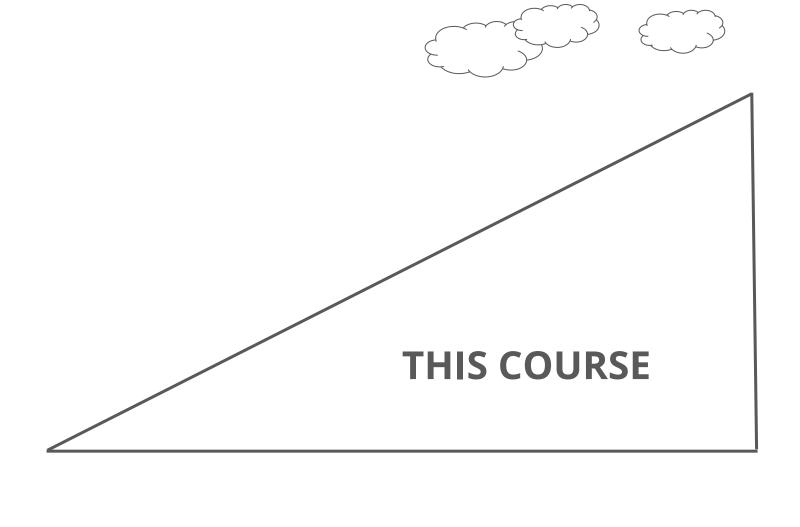
- 1. Problem framing → From a business problem to an ML system
- 2. ML System design
 - ◆ Feature-Training-Inference Pipelines
 - Infrastructure

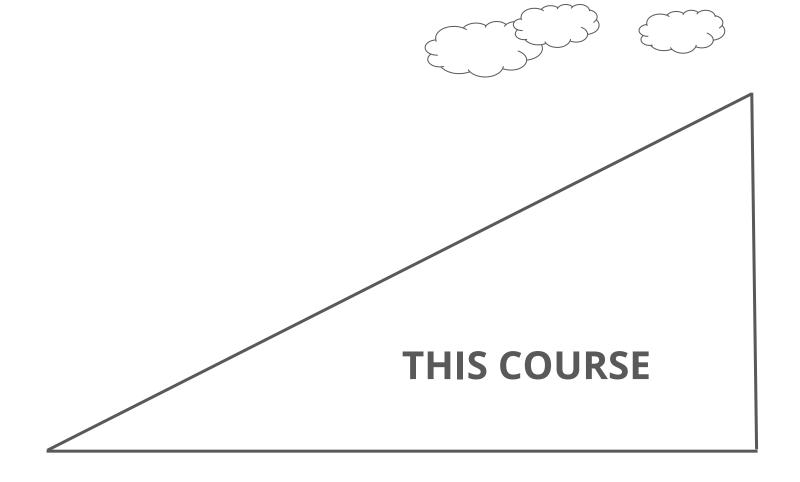
3. Start building Feature Pipeline

Are you ready to work hard ?



What is this course about ?









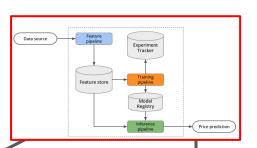
Training ML model from static data



THIS COURSE



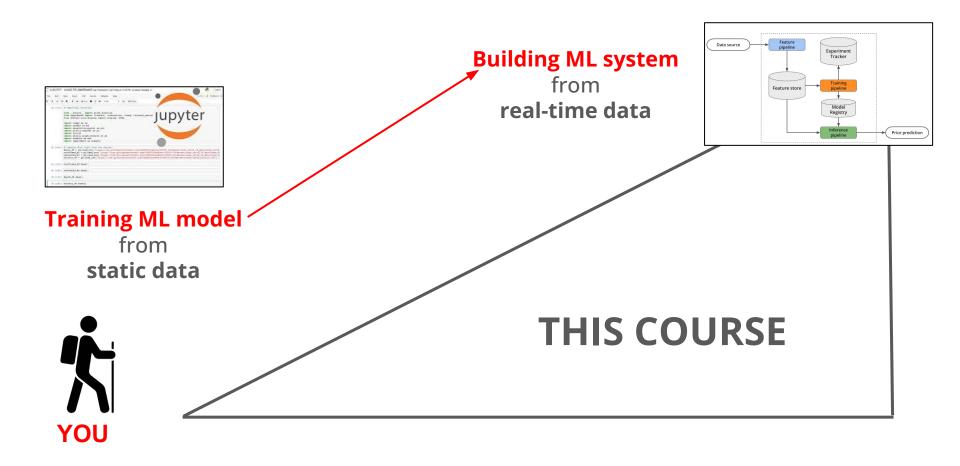
Building ML system from real-time data

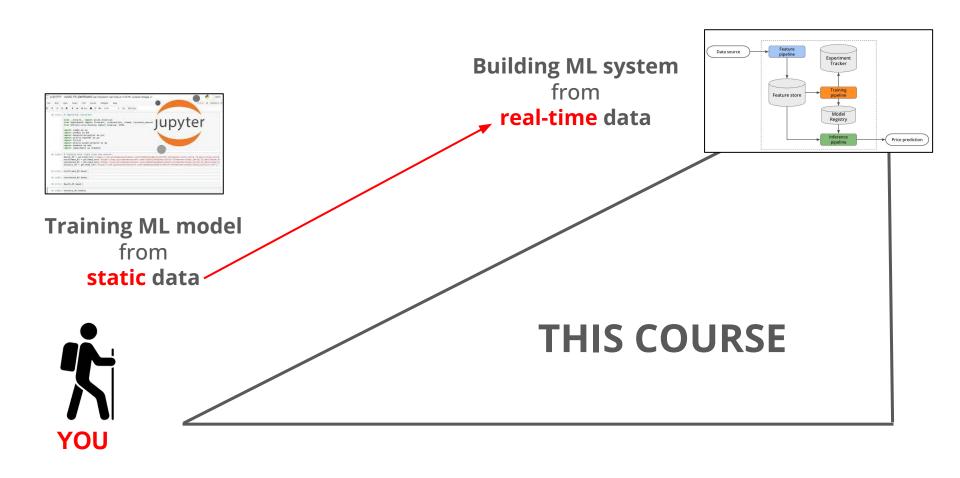


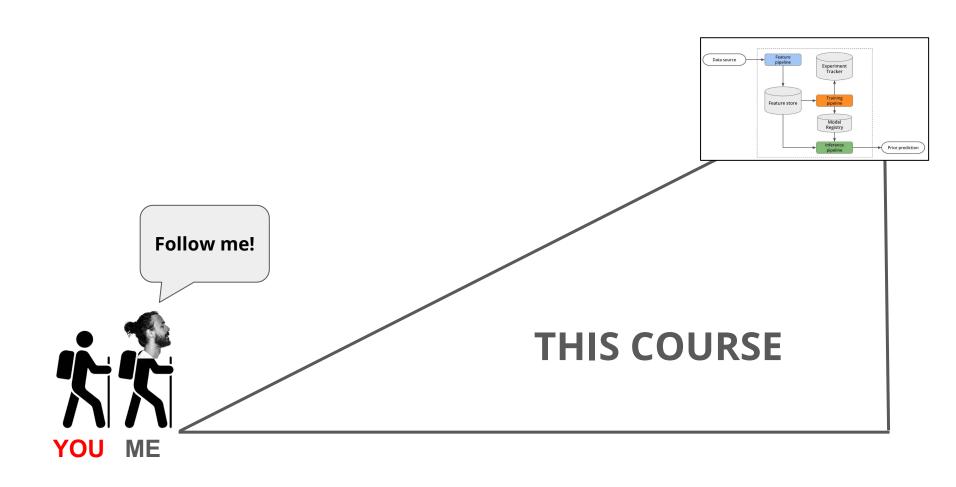
Training ML model from static data

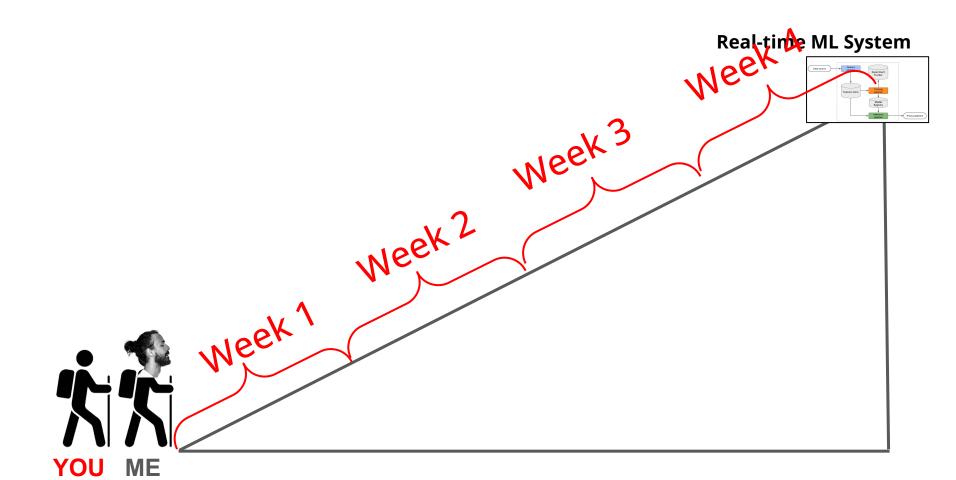


THIS COURSE





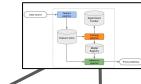




- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Comp
- > Linting and for thing
- > Feature Stores
- > Training to a generation
- > ML model trailing are experimentation
- > ML model registries
- > Real-time initerence with websocket
- > Deploy the whole thing
- > Monitoring of model and latency

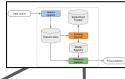


- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and from the
- > Feature Stor
- > Training data gent ation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



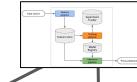
- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compo.
- > Linting and for king
- > Feature Stores
- > Training data gen ration
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency





- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and forma
- > Feature Stores
- > Training data genera ion
- > ML model training and overimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency





- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code

> Python Poetry

- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatt
- > Feature Stores
- > Training data generation
- > ML model training and perimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



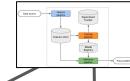


- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry

> Docker

- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation 00
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency





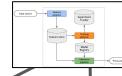
- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker

> Quix Streams

- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency







- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency







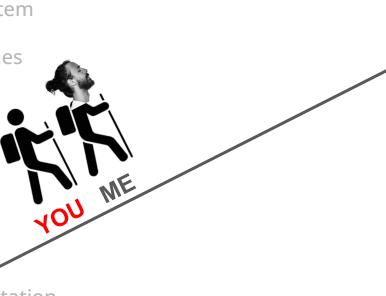
- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose

> Linting and formatting

- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting

> Feature Stores

- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



- > From business problem to ML syst
- > ML System design
- > Feature-Training-Inference pipeline
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting

> Feature Stores

- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



- > From business problem to ML syst
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting

> Feature Stores

- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



- > From business problem to ML syst
- > ML System design
- > Feature-Training-Inference pipeline
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting

> Feature Stores

- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



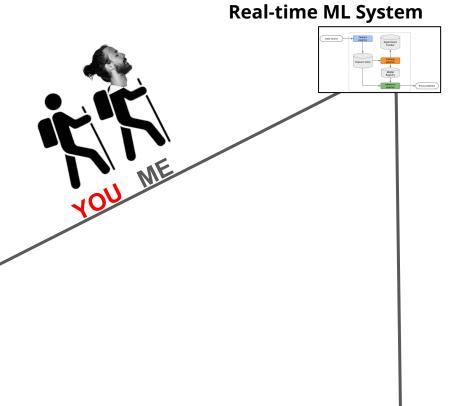


> From business problem to ML system

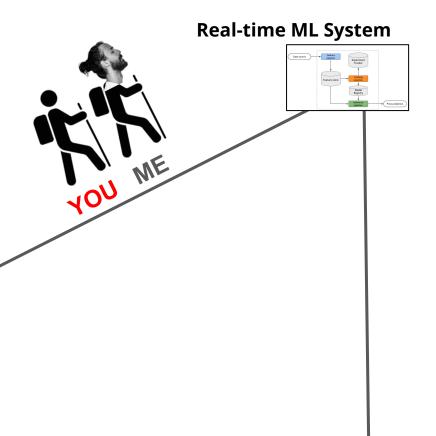
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



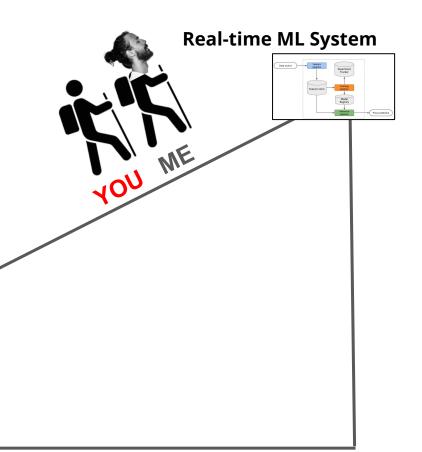
- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries .
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency

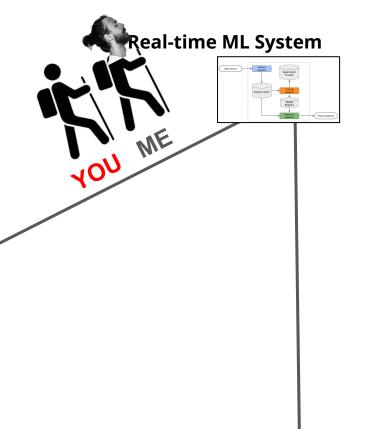


- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



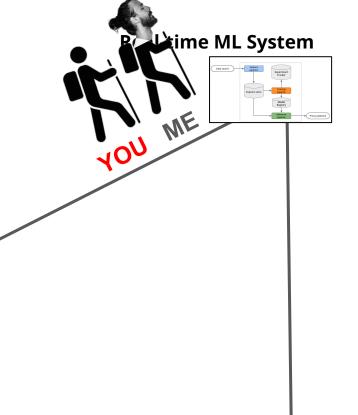
This is what you will learn

- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



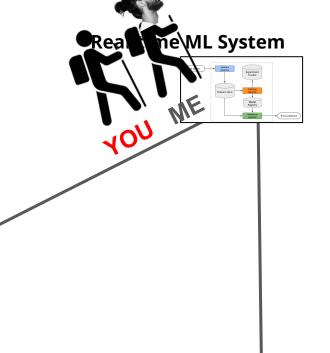
This is what you will learn

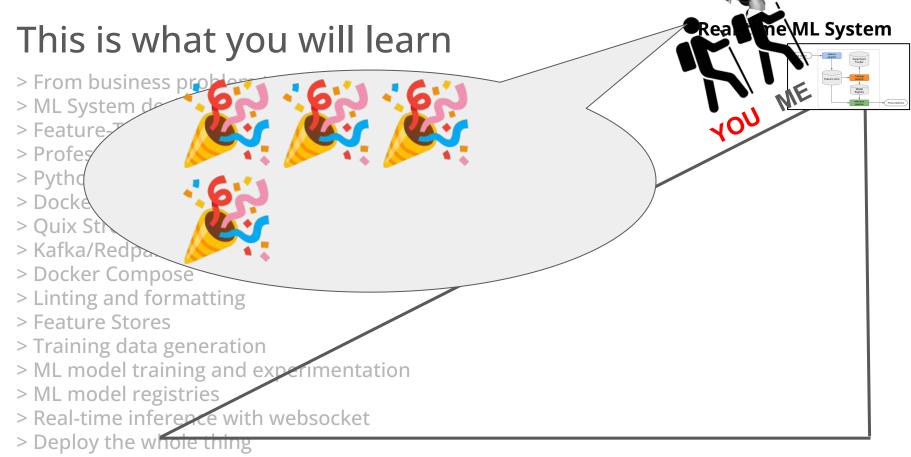
- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency



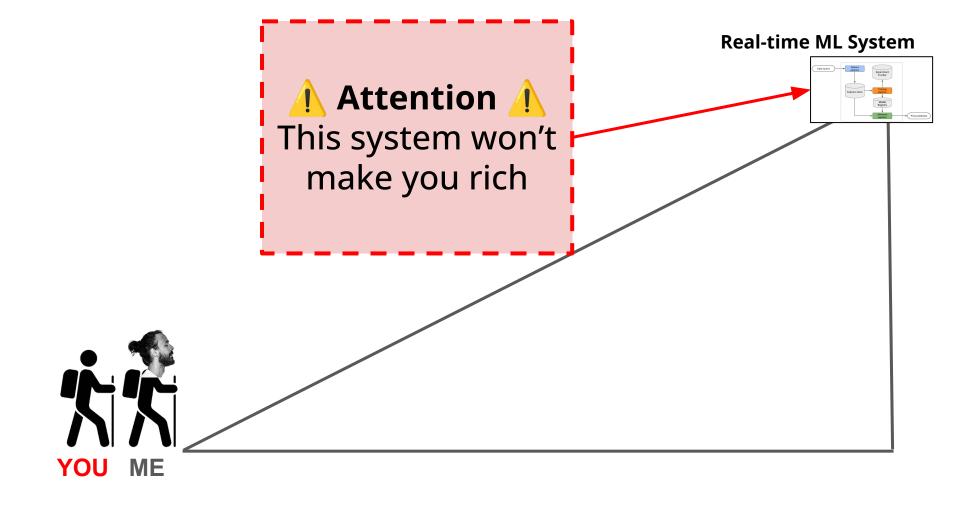
This is what you will learn

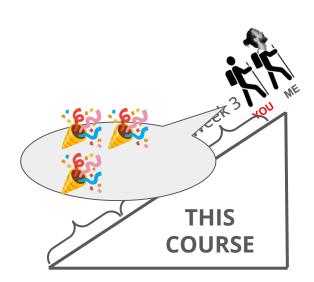
- > From business problem to ML system
- > ML System design
- > Feature-Training-Inference pipelines
- > Professional Python code
- > Python Poetry
- > Docker
- > Quix Streams
- > Kafka/Redpanda
- > Docker Compose
- > Linting and formatting
- > Feature Stores
- > Training data generation
- > ML model training and experimentation
- > ML model registries
- > Real-time inference with websocket
- > Deploy the whole thing
- > Monitoring of model and latency





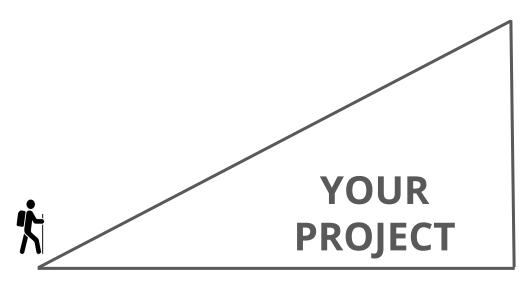
> Monitoring of model and latency

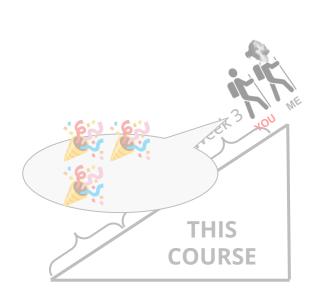


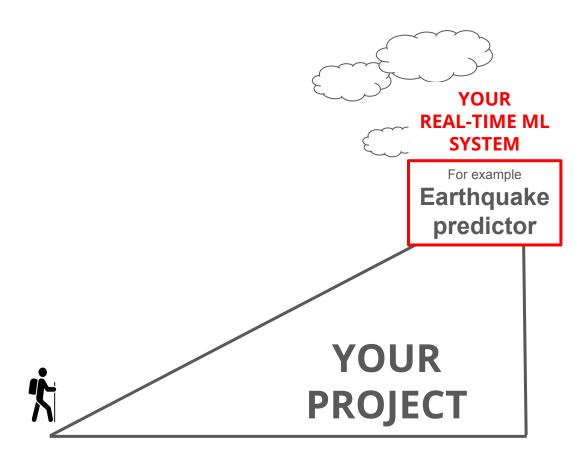


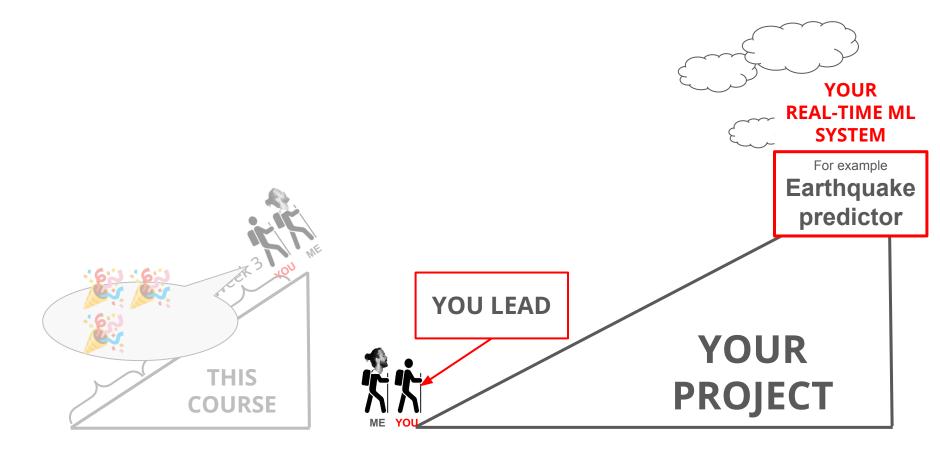


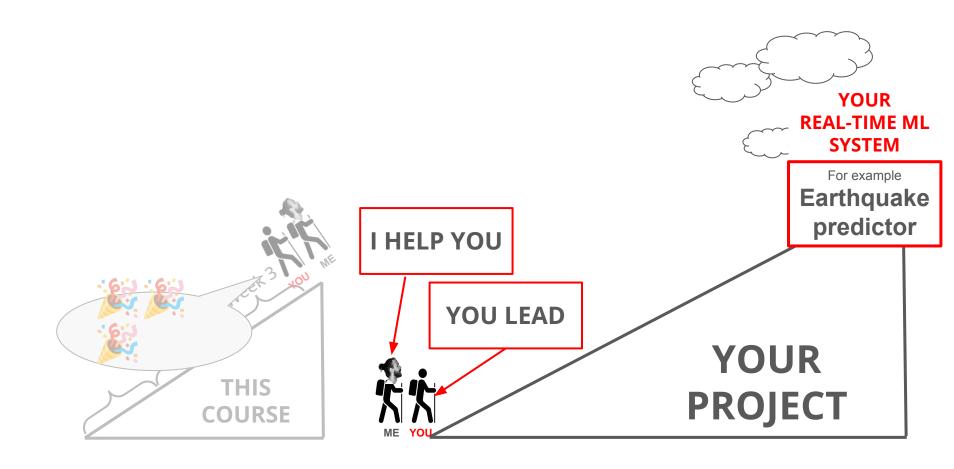




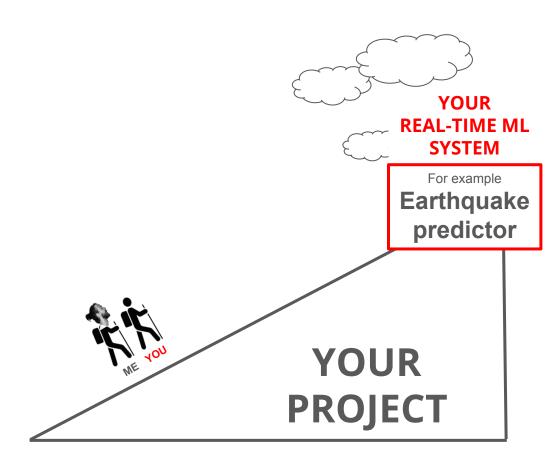


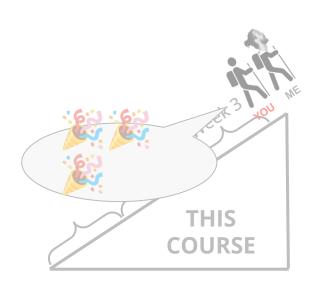


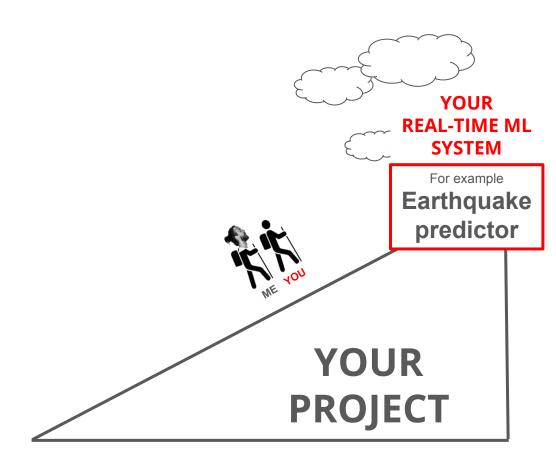




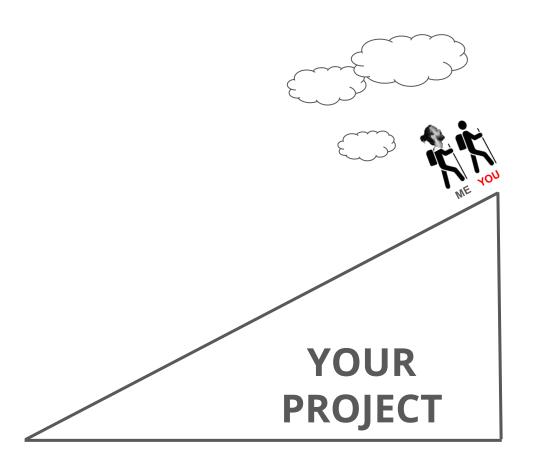


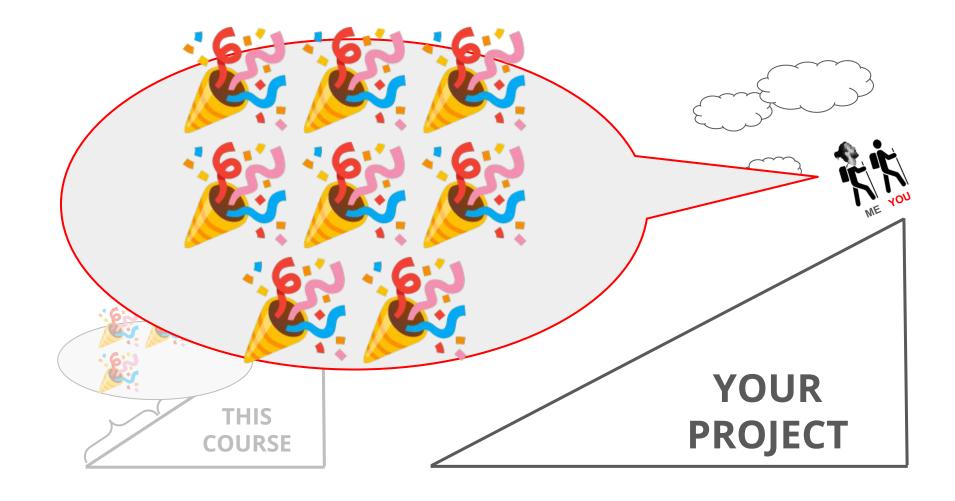




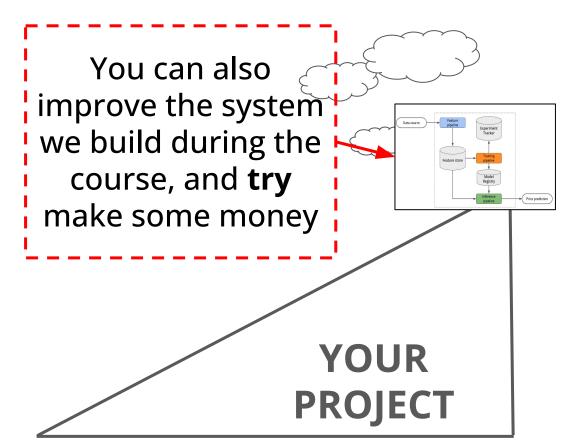












Problem framing

From a Business Problem to an ML System

The Business Problem

→ We work in a trading firm, and want to increase the daily profit of our trading bot.

1. The Business metric we want to impact

→ We work in a trading firm, and want to increase the daily profit of our trading bot.
 1. Busines

1. Business metric (KPI)

2. The End User of our solution

→ We work in a trading firm, and want to increase the daily profit of our trading bot.
 1. Business metric (KPI)
 2. End user

3. The ML Proxy metric

→ We work in a trading firm, and want to increase the daily profit of our trading bot.
 → Let's build a system to predict
 1. Business metric (KPI)
 2. End user

→ Let's build a system to predict short-term crypto prices

3. The ML Proxy metric we will predict

→ We work in a trading firm, and want to increase the daily profit of our trading bot.
 → Let's build a system to predict short-term crypto prices
 → 3. The ML Proxy metric

3. The ML Proxy metric we will predict

→ We work in a trading firm, and want to increase the daily profit of our 1. Business metric (KPI) trading bot. 2. End user Let's build a system to predict short-term crypto prices -3. The ML Proxy metric For example **9** A system that predicts the price of Ethereum in the **next 10 seconds**.

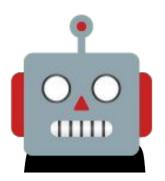
4. Data Sources

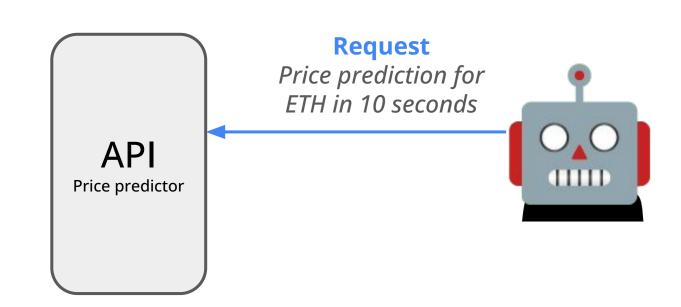
→ We work in a trading firm, and want to increase the daily profit of our trading bot.
 → Let's build a system to predict short-term crypto prices using real-time market data.
 1. Business metric (KPI)
 2. End user
 3. The ML Proxy metric

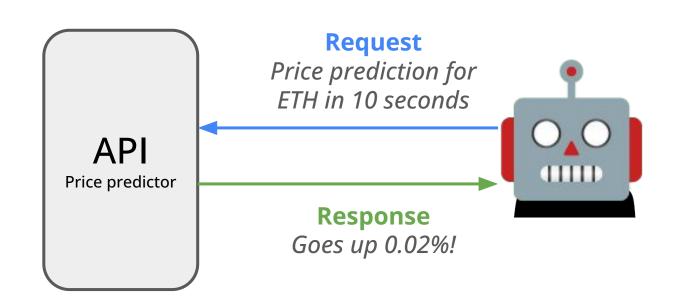
4. Data Sources

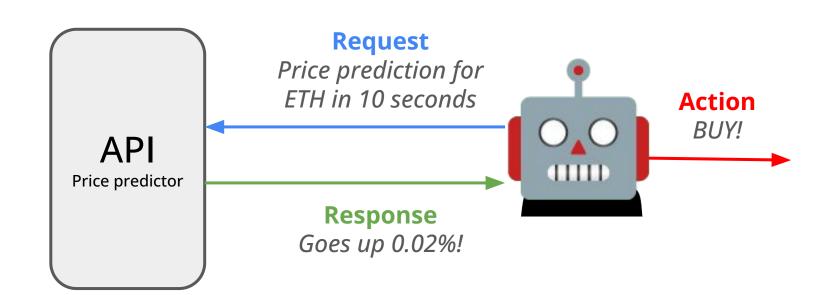
→ We work in a trading firm, and want to increase the daily profit of our trading bot.
 → Let's build a system to predict short-term crypto prices using real-time market data.
 → 4. Data sources





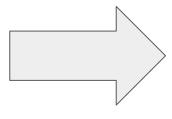






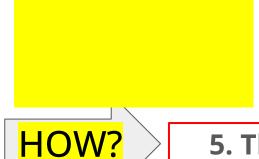
- 1. Business metric (KPI)
- 2. End user
- 3. The ML Proxy metric
- 4. Data sources

- 1. Business metric (KPI)
- 2. End user
- 3. The ML Proxy metric
- 4. Data sources

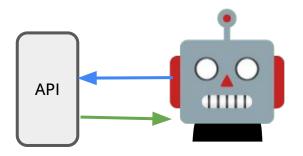


5. The ML System

- 1. Business metric (KPI)
- 2. End user
- 3. The ML Proxy metric
- 4. Data sources



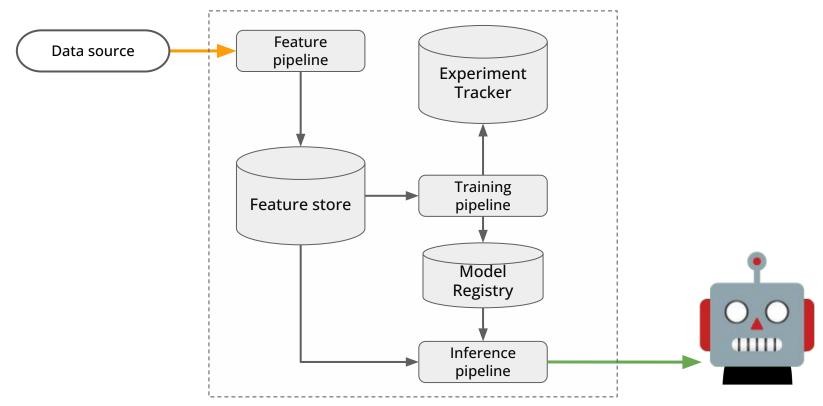
5. The ML System



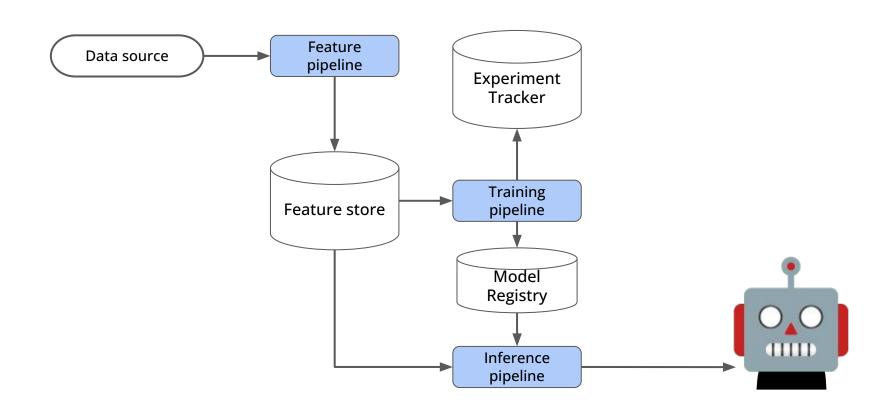
The **ML system** behind our API From raw market data to real-time predictions



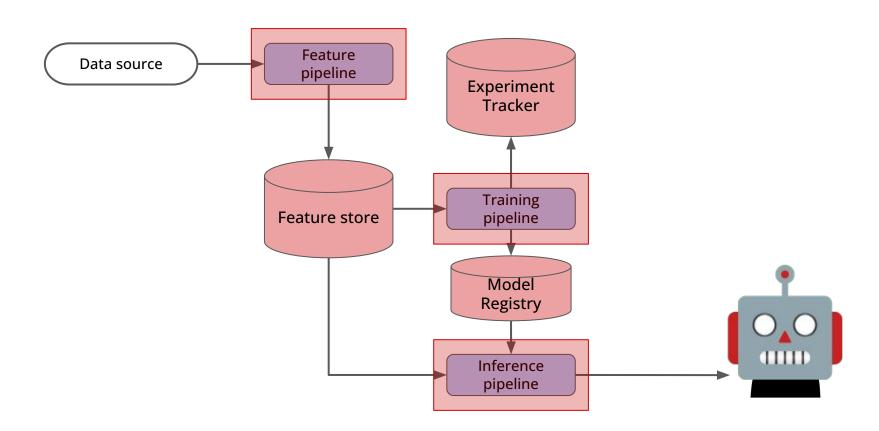
The **ML system** behind our API From raw market data to real-time predictions



ML system = Pipelines



ML system = Pipelines + Infrastructure



What is a Pipeline?

A Pipeline

→ is a computer program (e.g Python script)

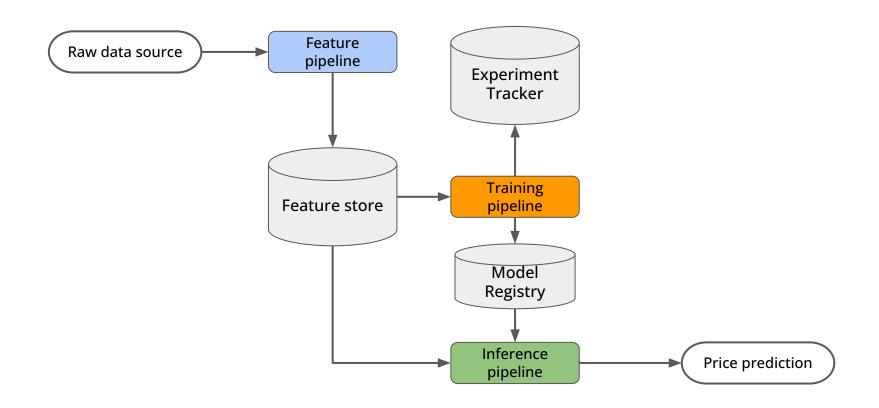
A Pipeline

- → is a computer program (e.g Python script)
- → with clearly defined **inputs** and **outputs**

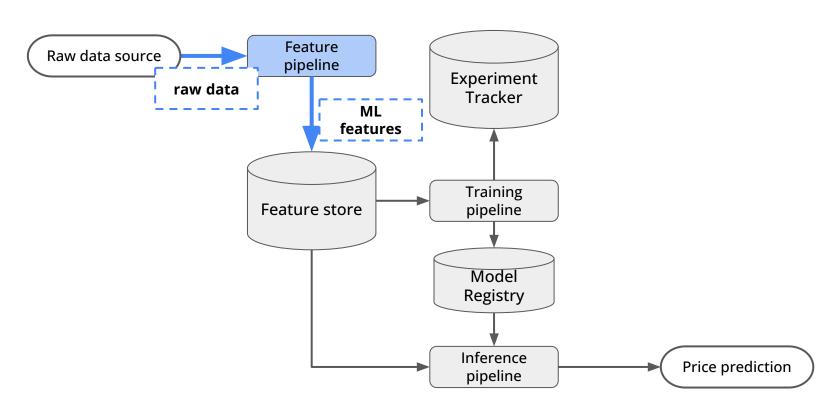
A Pipeline

- → is a computer program
- → with clearly defined inputs and outputs
- → that runs on a **schedule** or **continuously**

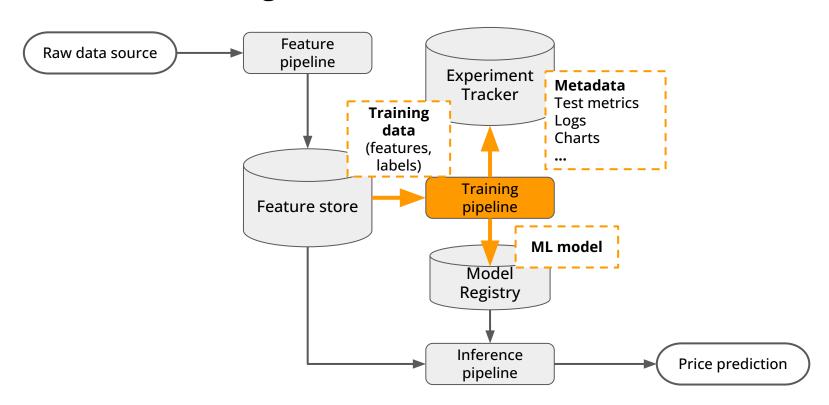
Feature-Training-Inference pipelines



1. Feature pipeline From raw data to reusable features for ML models

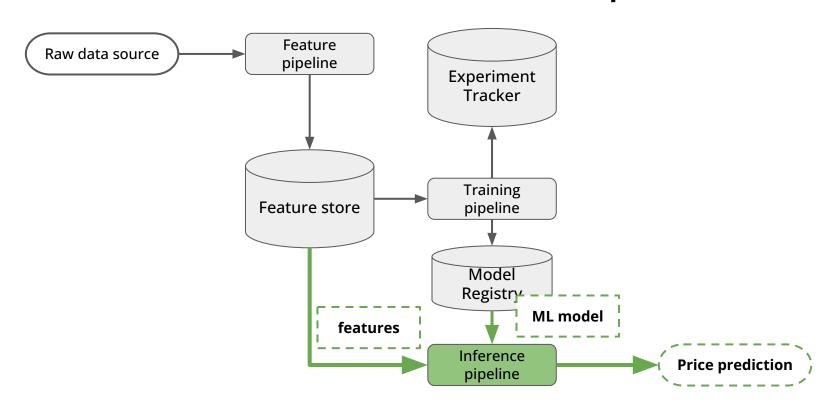


2. Training pipeline From training data to an ML model + metadata

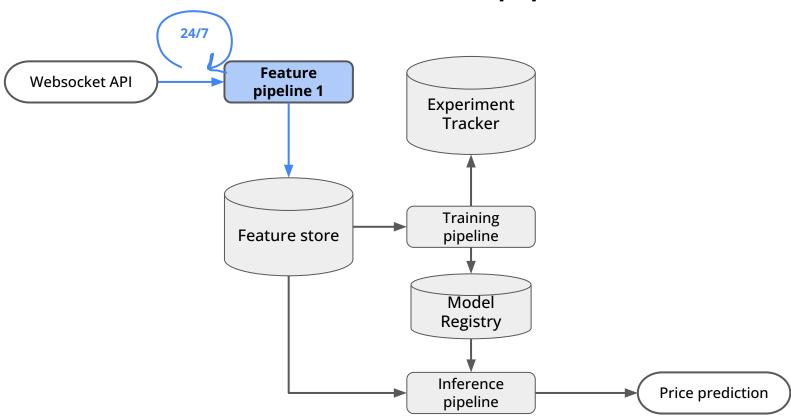


3. Inference pipeline

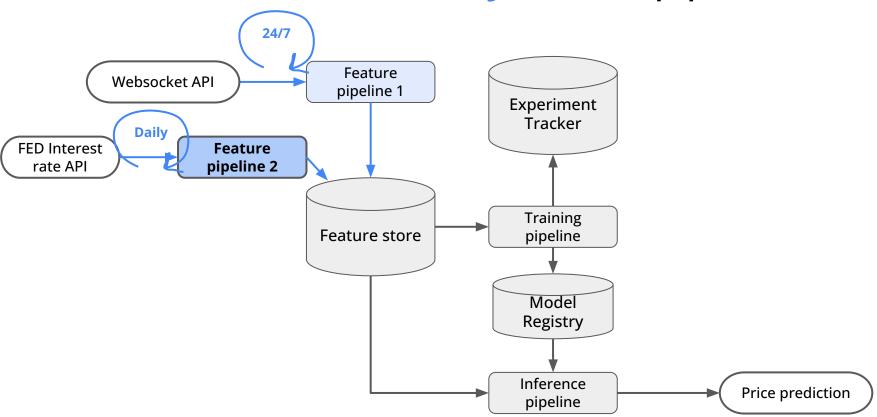
From an ML model and fresh features to predictions



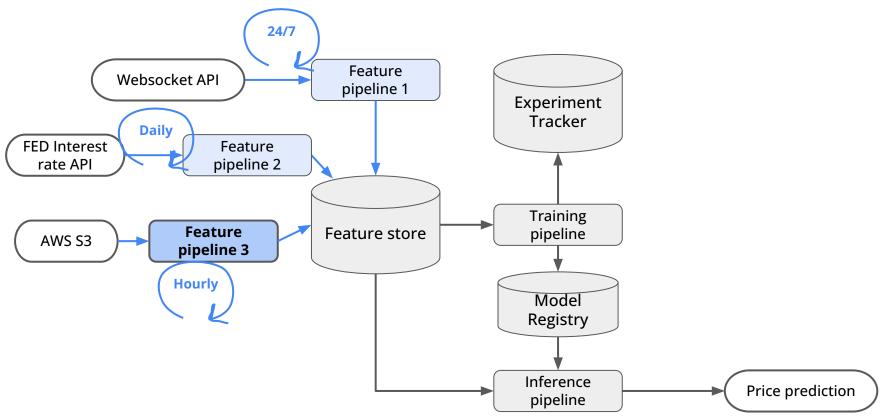
1 real-time feature pipeline



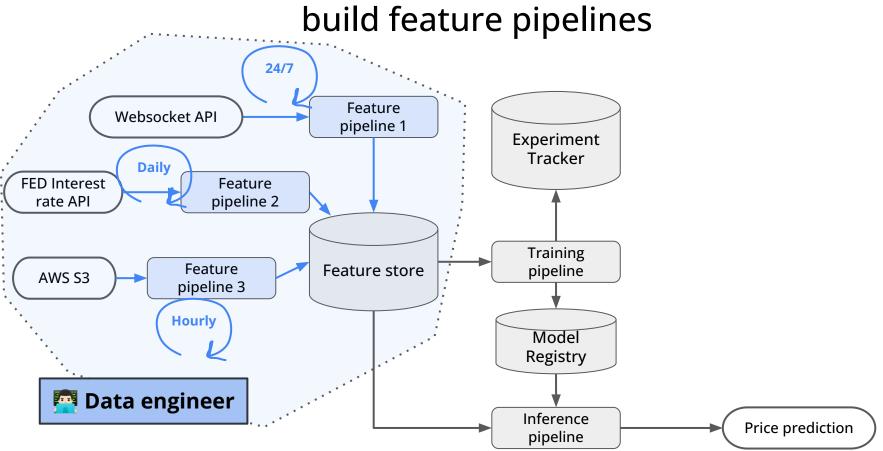
1 real-time + 1 daily feature pipeline



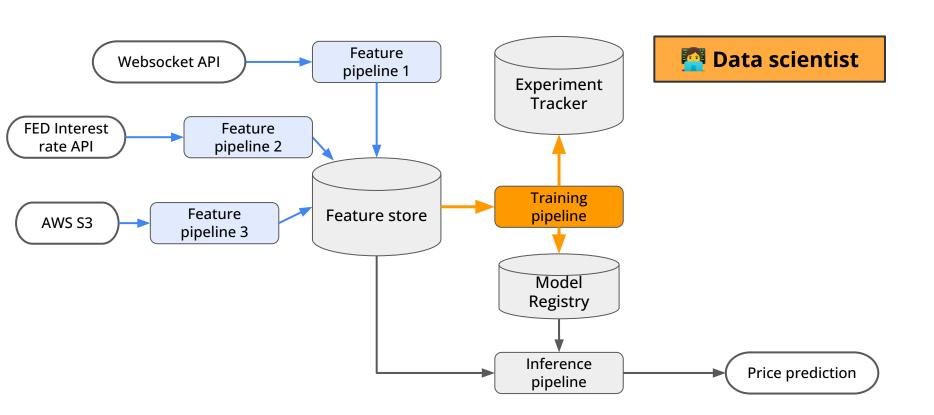
1 real-time + 1 daily + 1 hourly feature pipeline



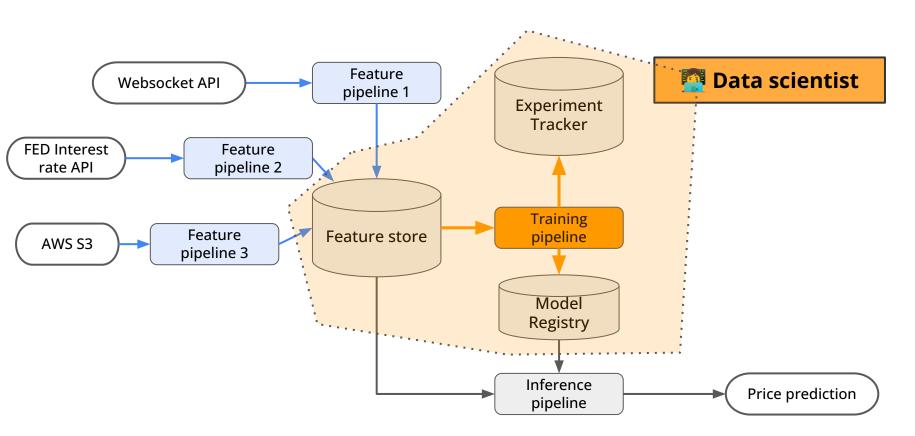
Data engineers



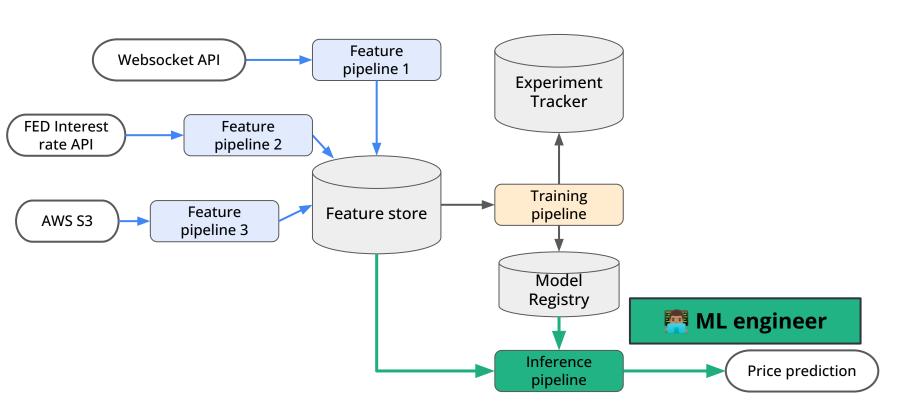
Data scientists train ML models



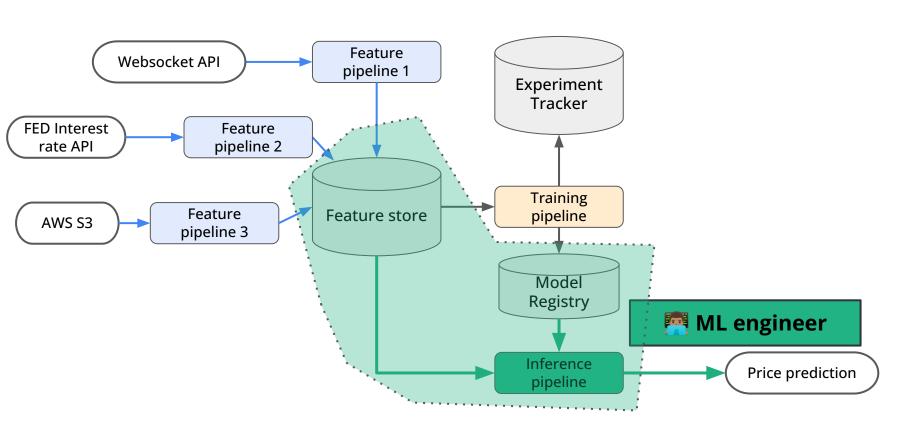
Data scientists train ML models



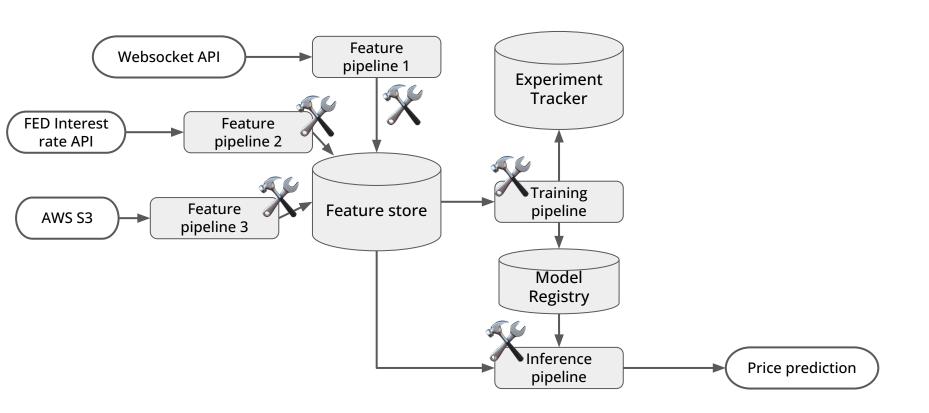
ML engineers deploy these models



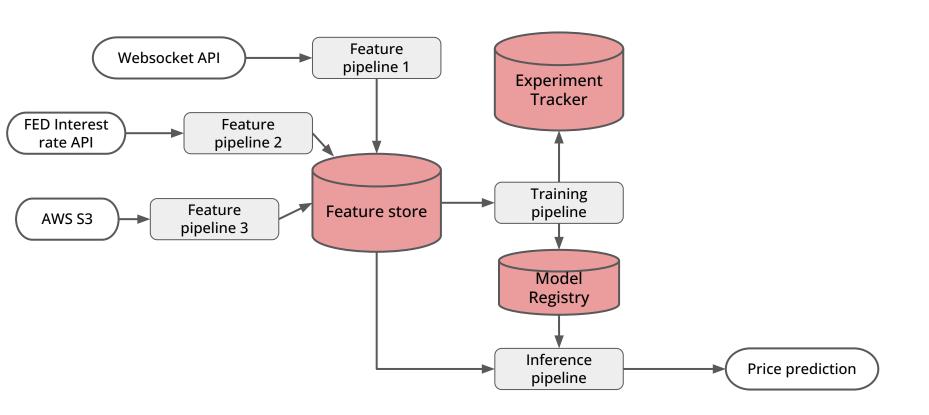
ML engineers deploy these models



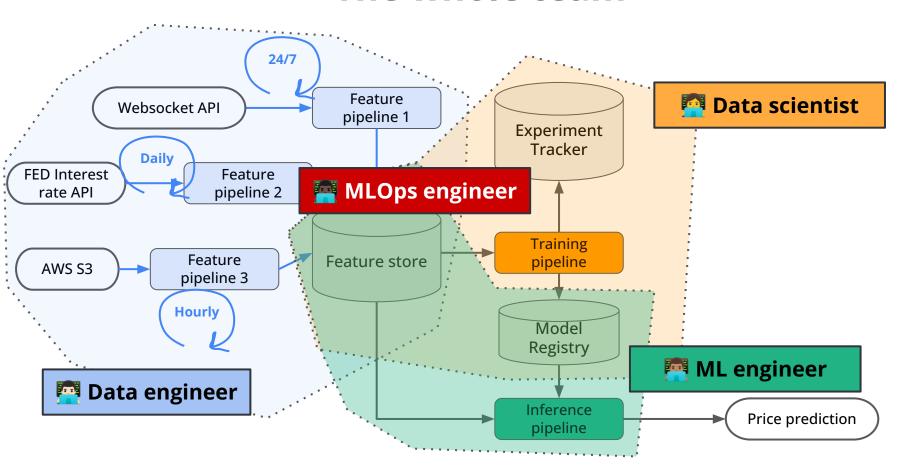
ML Ops engineers Build tools for the rest of the team



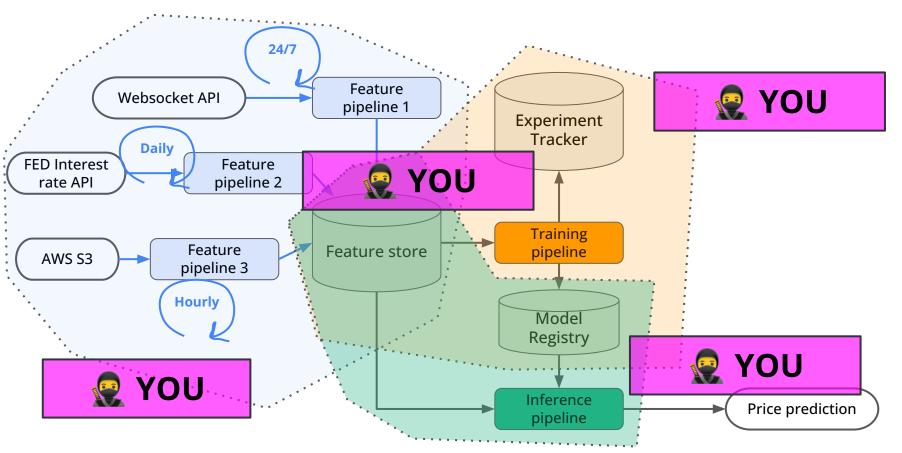
ML Ops engineers Take care of infrastructure



The whole team



ML ninjas do everything 🥷



But HOW?

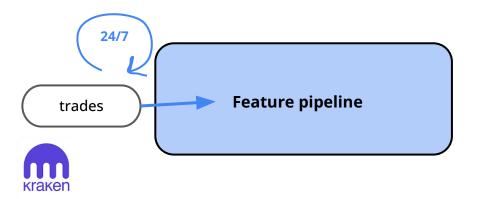
System Design



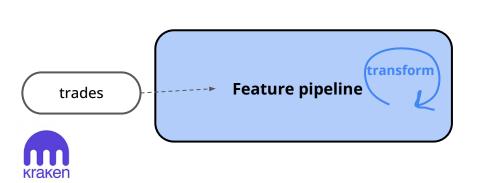
Feature pipeline

Feature pipeline

Real-time Feature pipeline Reads stream of trades from Kraken Websocket

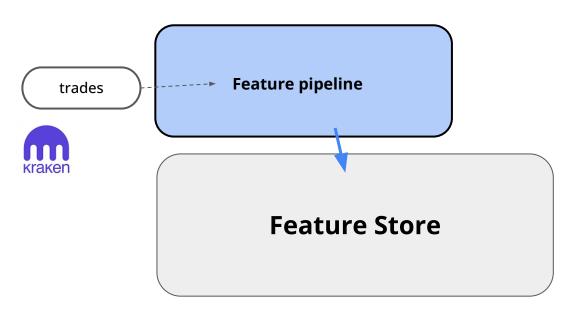


Real-time Feature pipeline Transforms them into **1-minute OHLC candles**

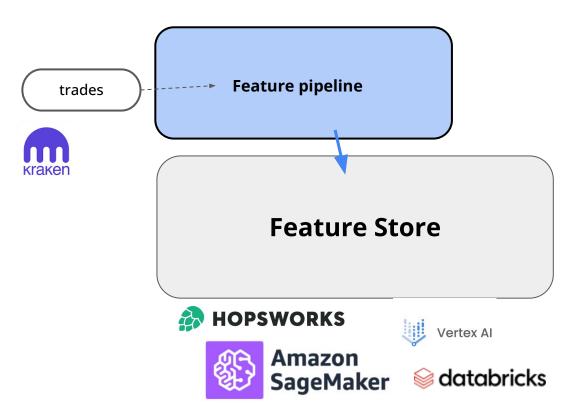




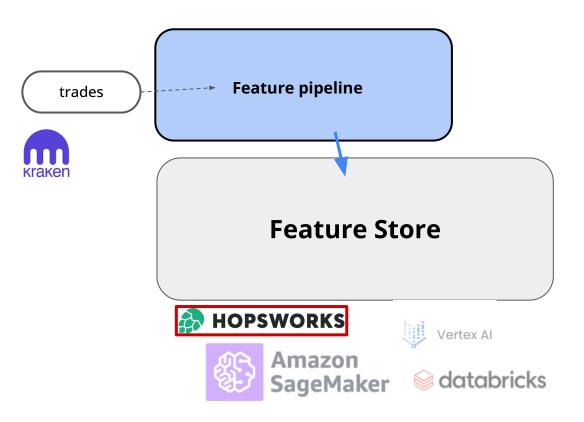
Real-time Feature pipeline Save 1-minute OHLC candles to **Feature Store**



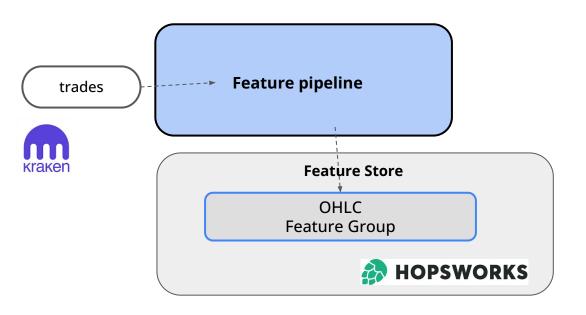
Real-time Feature pipeline Save 1-minute OHLC candles to **Feature Store**



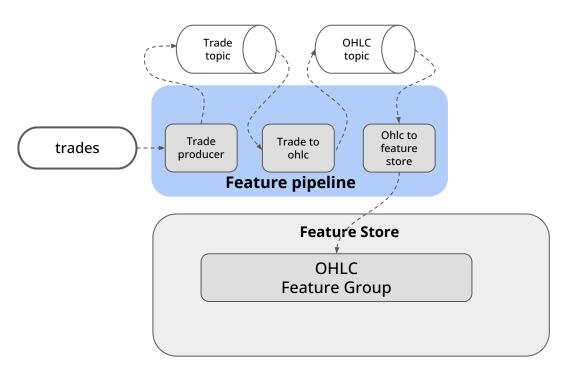
Real-time Feature pipeline Save 1-minute OHLC candles to **Feature Store**



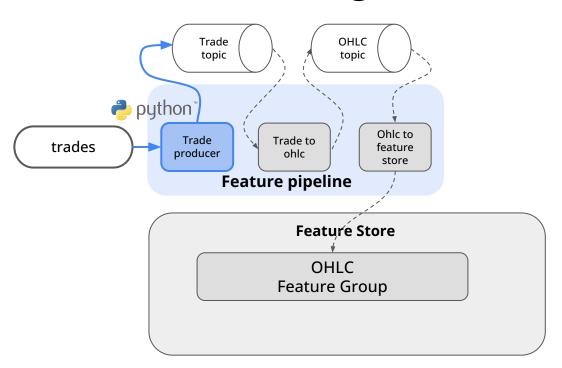
Real-time Feature pipeline More precisely, into a Feature Group



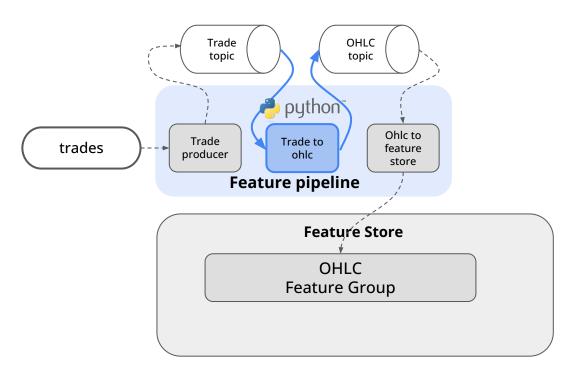
Modular Real-time Feature pipeline



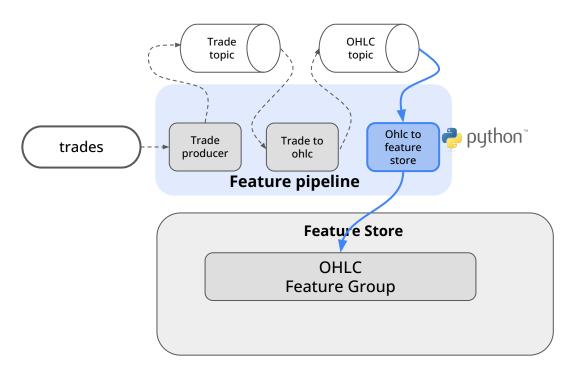
Ingests raw trades



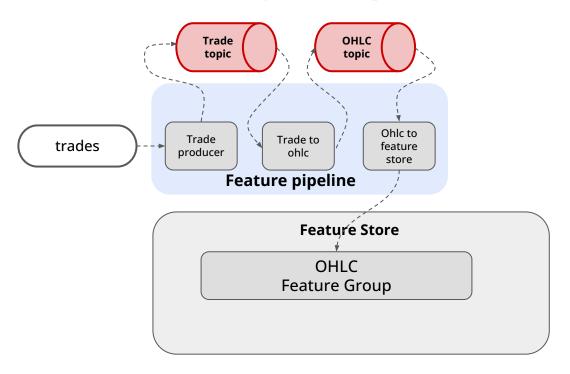
Transform raw trades into OHLC candles



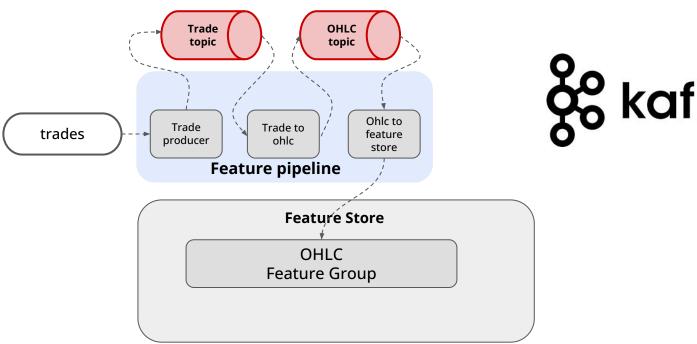
Pushed OHLC candles to the Feature Store



Streaming data platform for communication

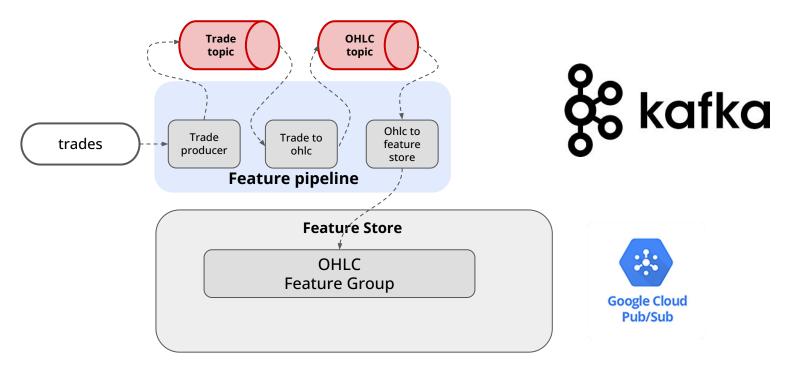


Streaming data platform for communication

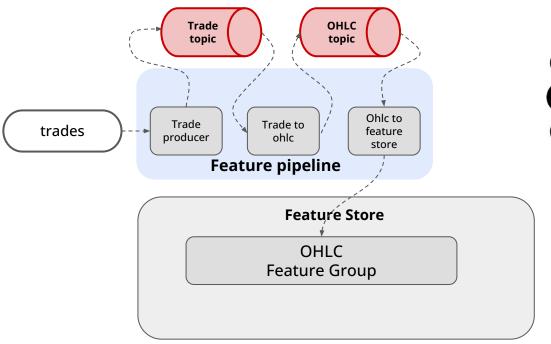




Streaming data platform for communication



Streaming data platform for communication

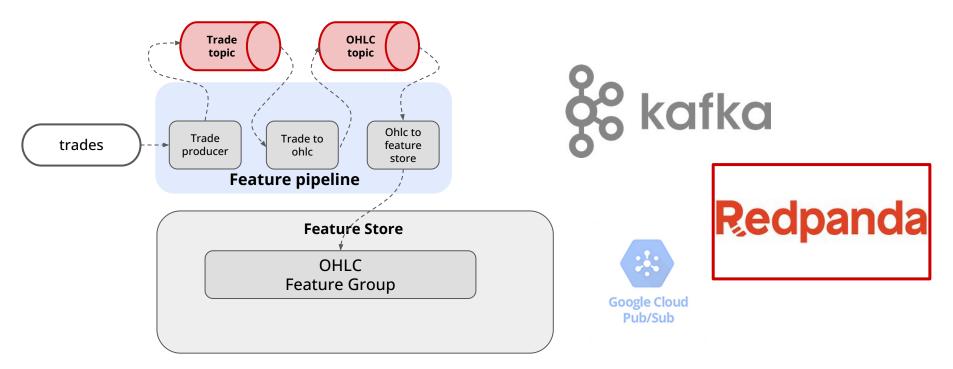




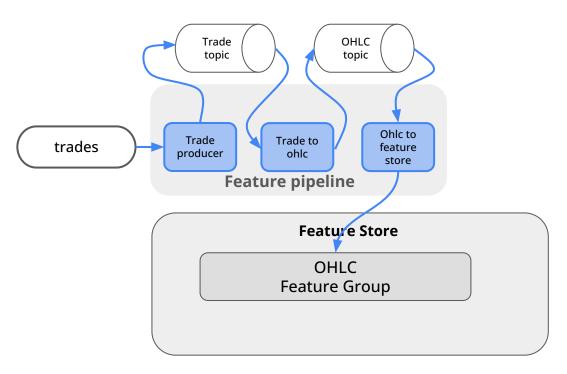




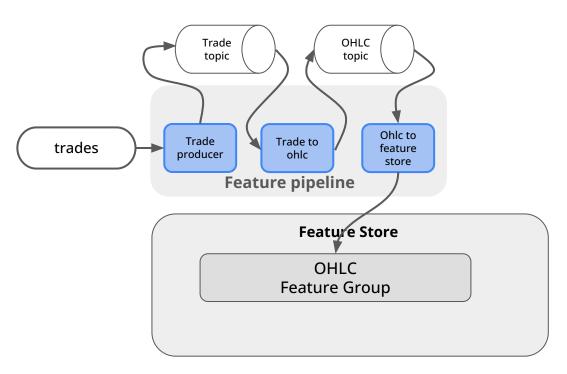
Streaming data platform for communication



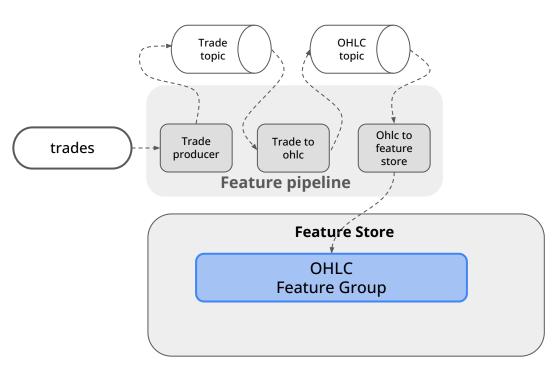
Backfill historical features for last 90 days



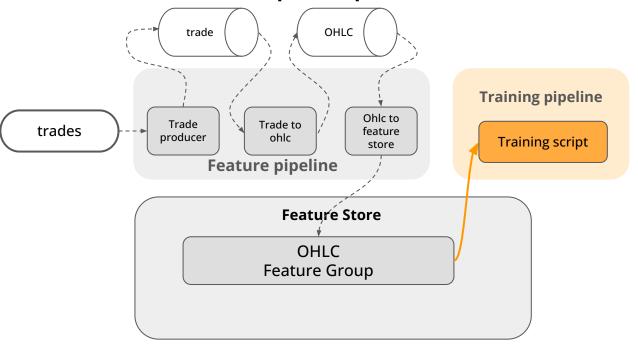
Backfill historical features for last 90 days



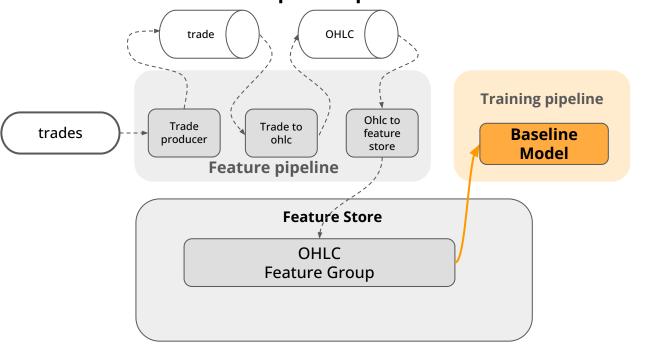
This is our training data



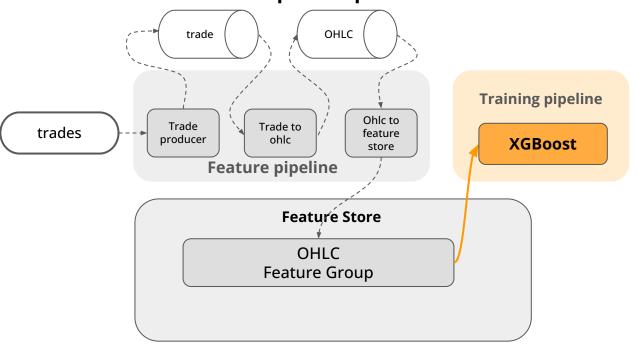
Build ML model from this training data



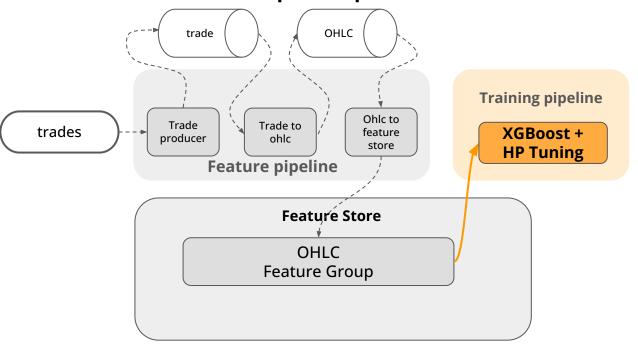
Model baseline



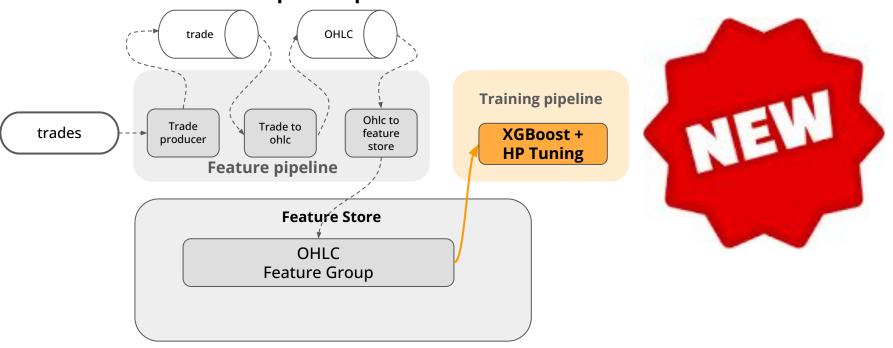
XGBoost

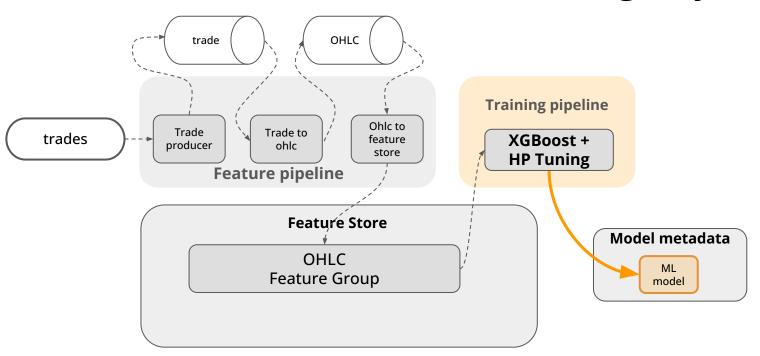


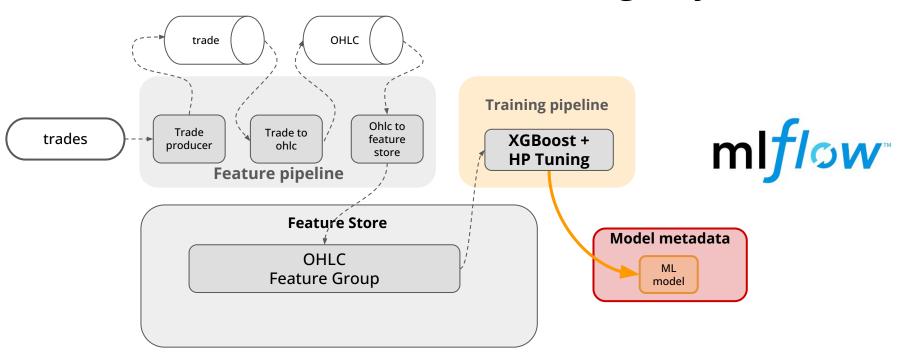
XGBoost + Hyperparameter Tuning

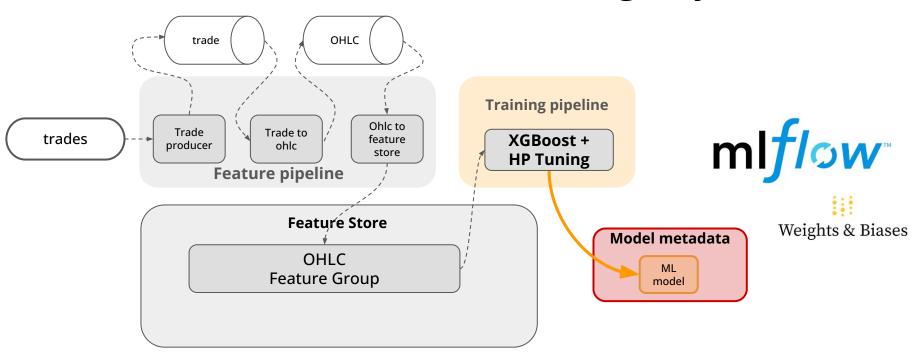


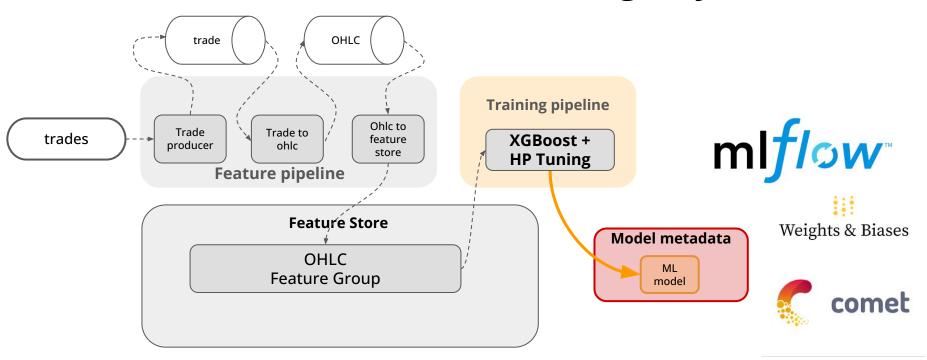
XGBoost + Hyperparameter Tuning

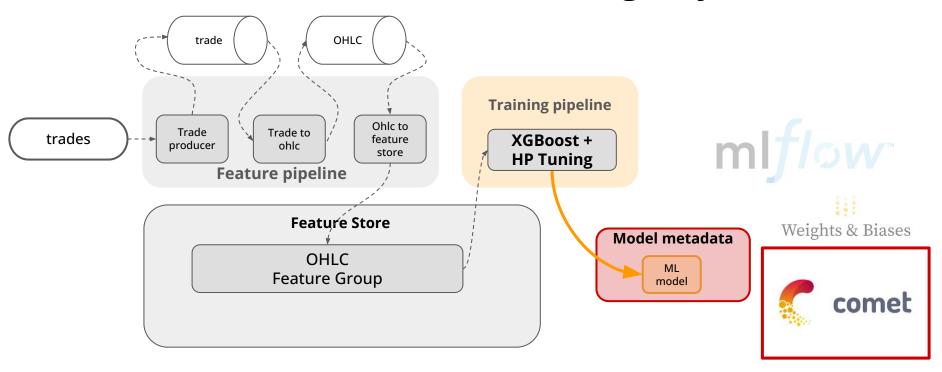




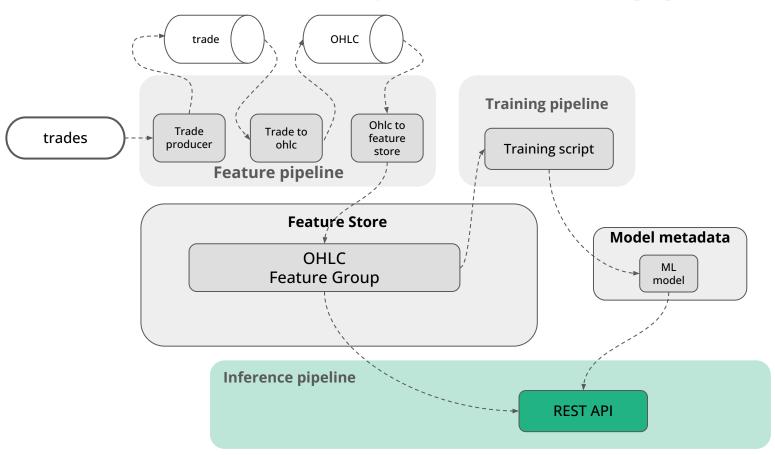




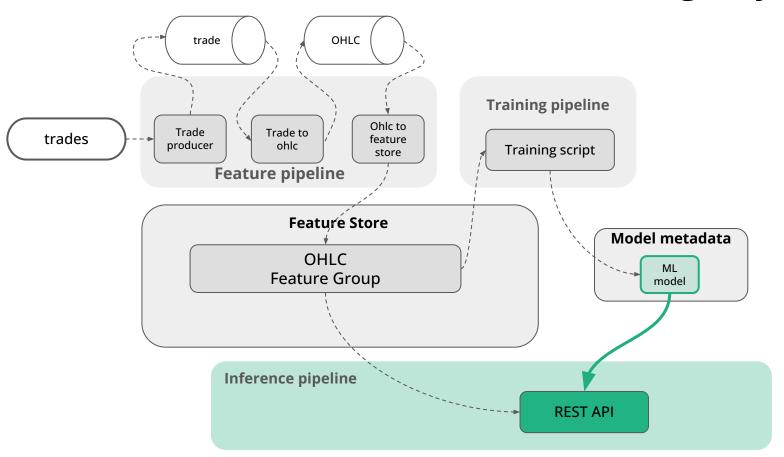




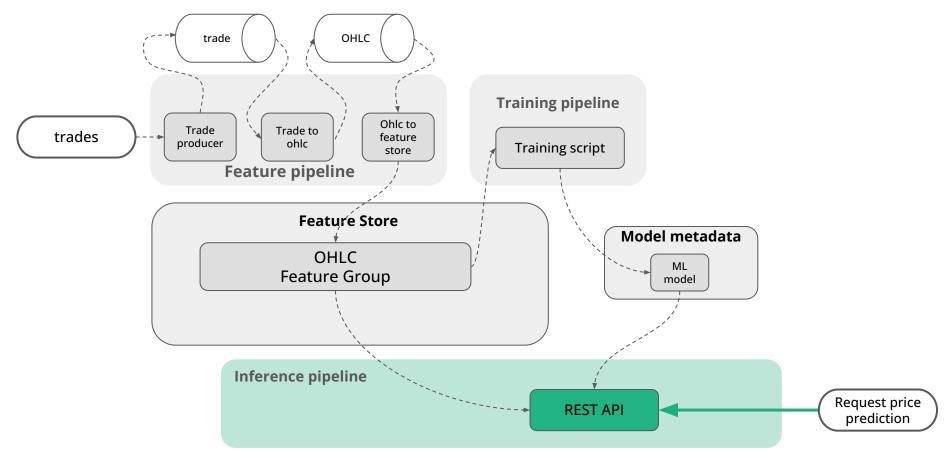
Start building the Inference pipeline



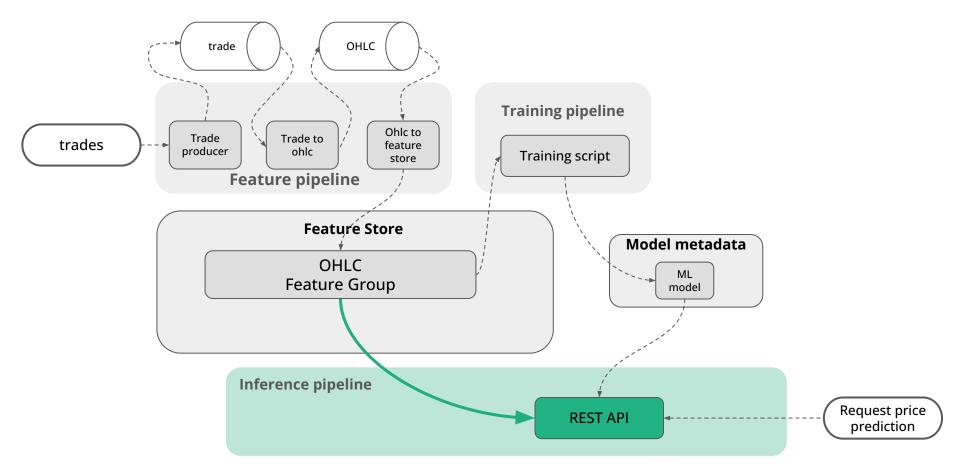
Load Production model from registry

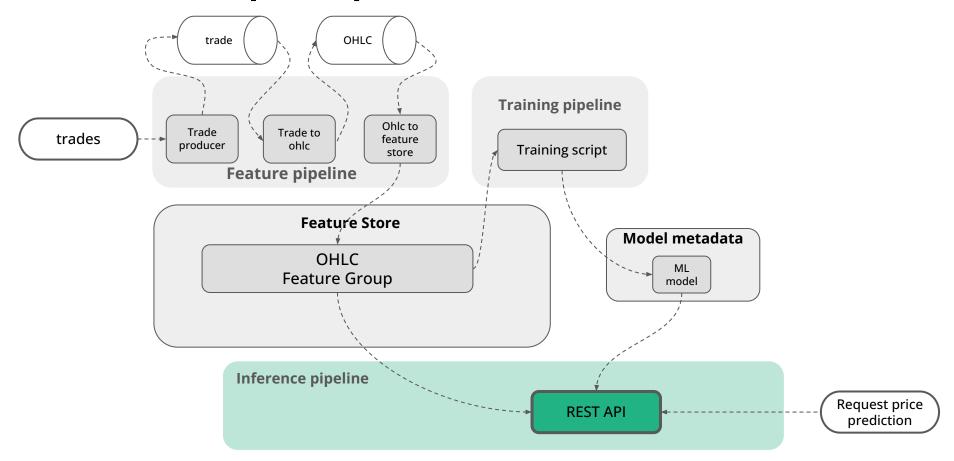


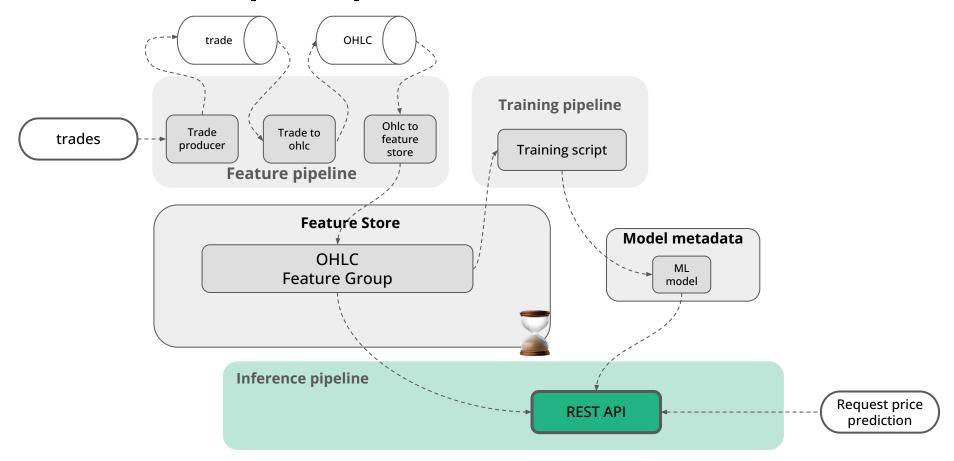
Listens to incoming requests

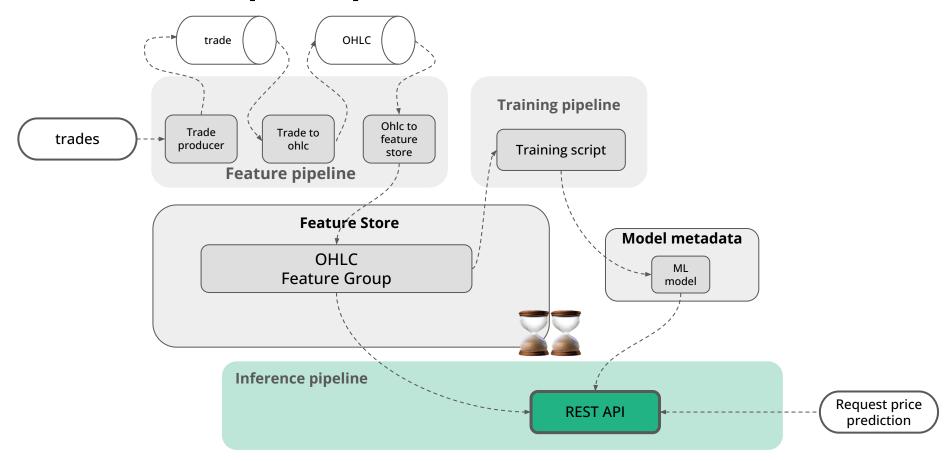


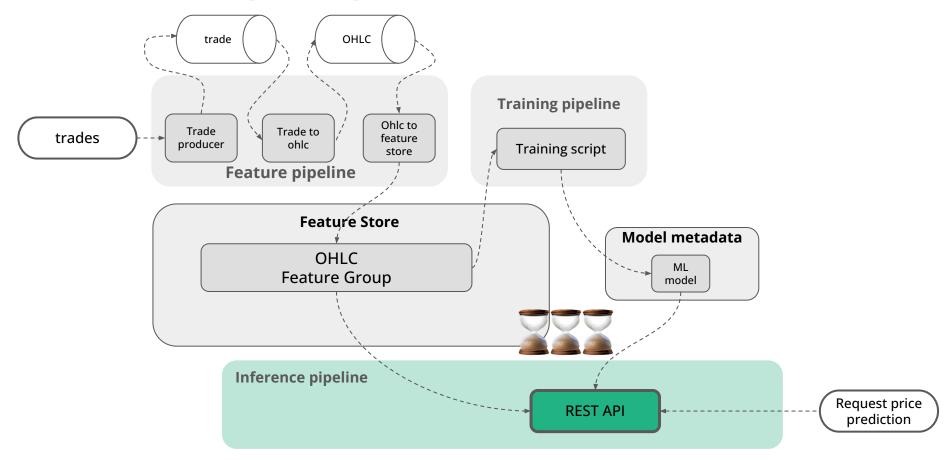
Gets fresh features from the store

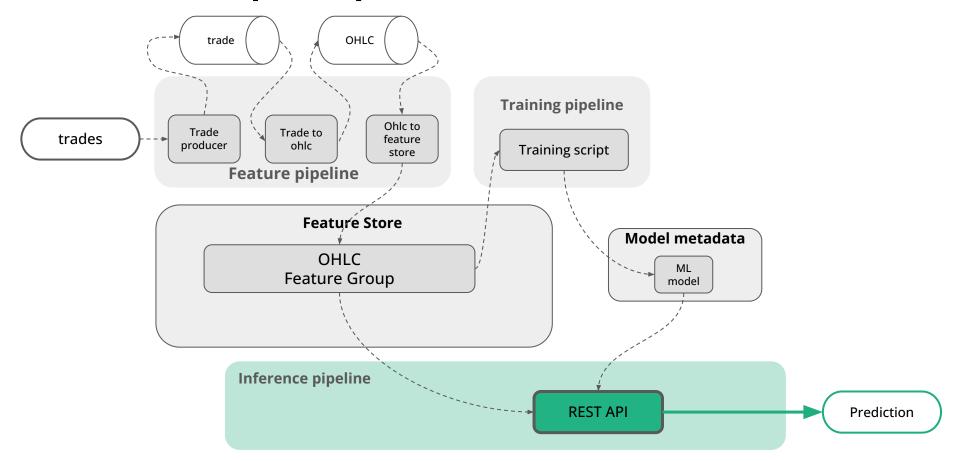








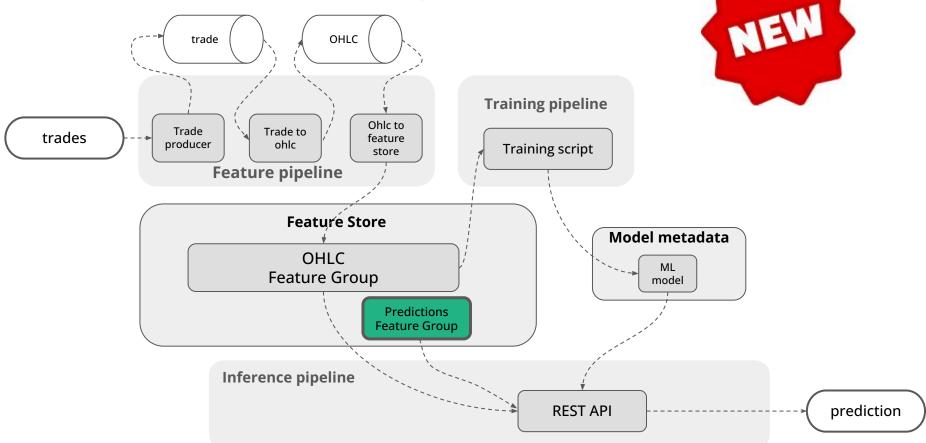




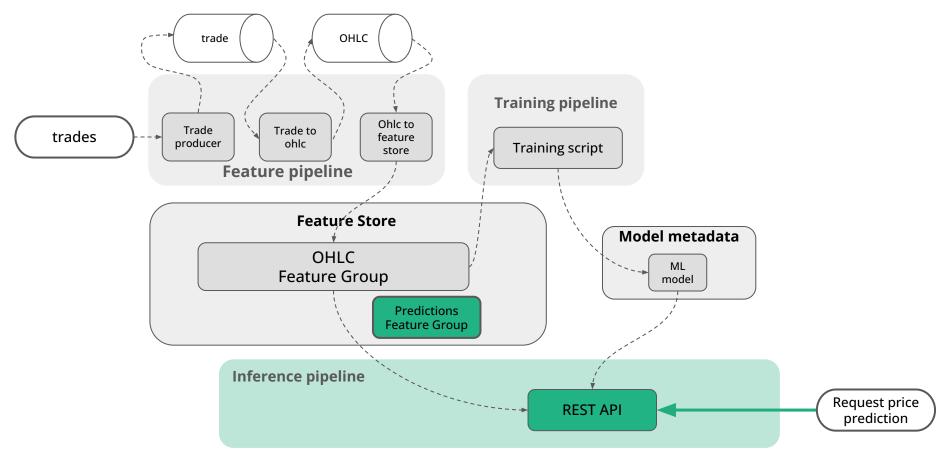


REST API caching

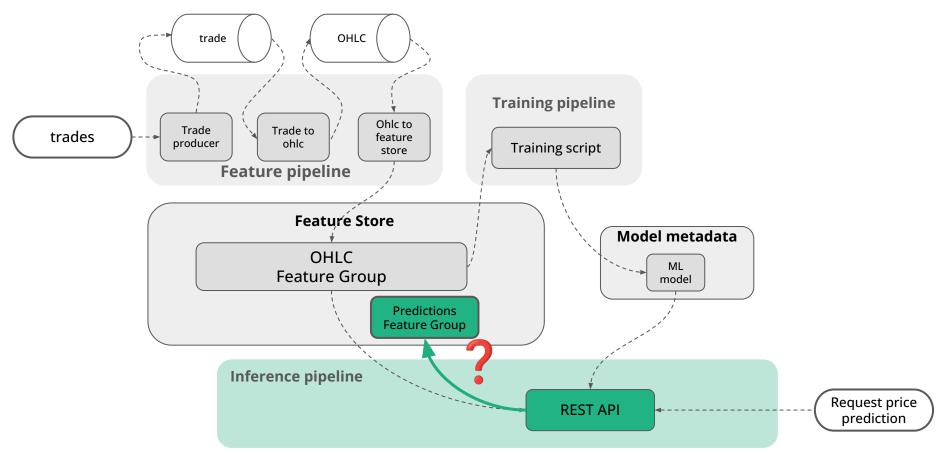
Add caching to our REST API



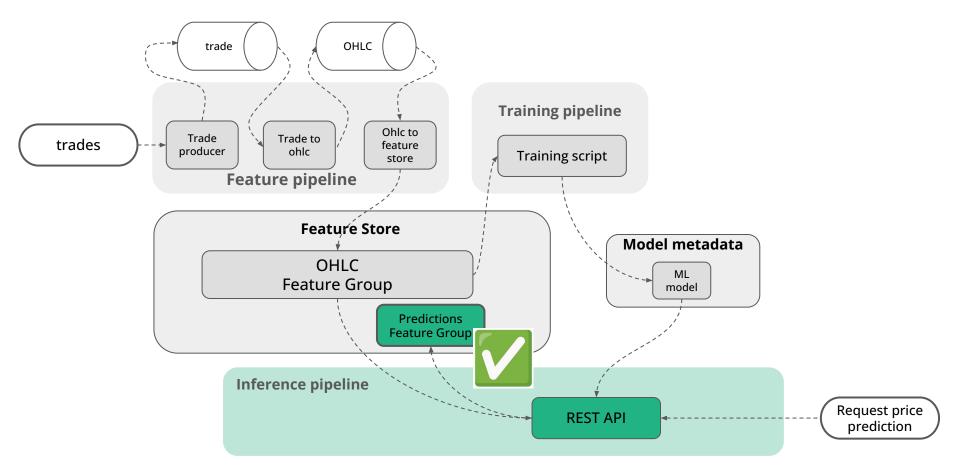
Listens to incoming requests



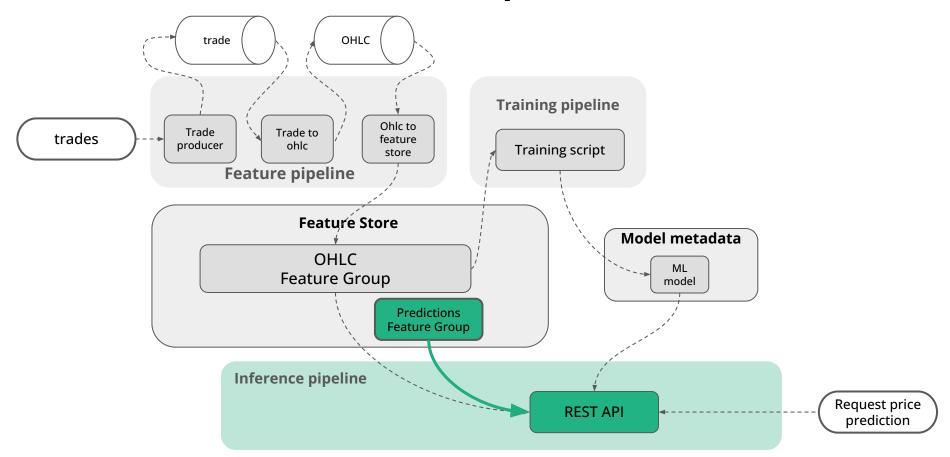
Checks if request is cached



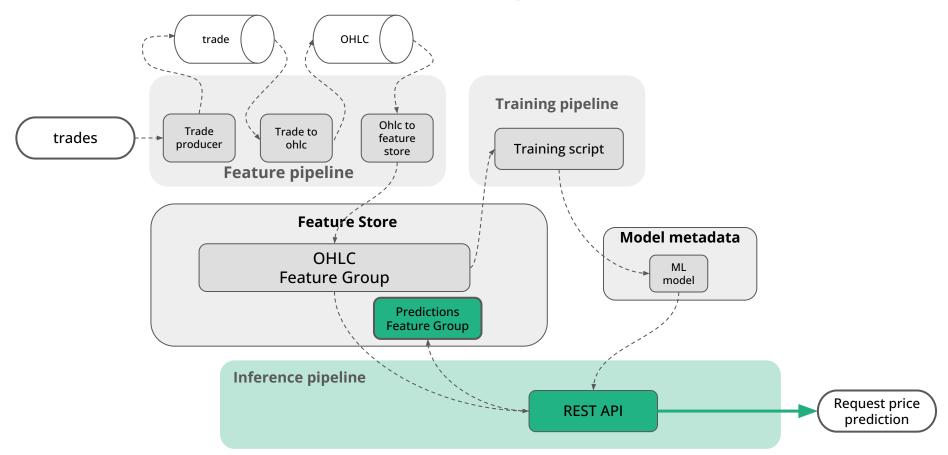
If YES



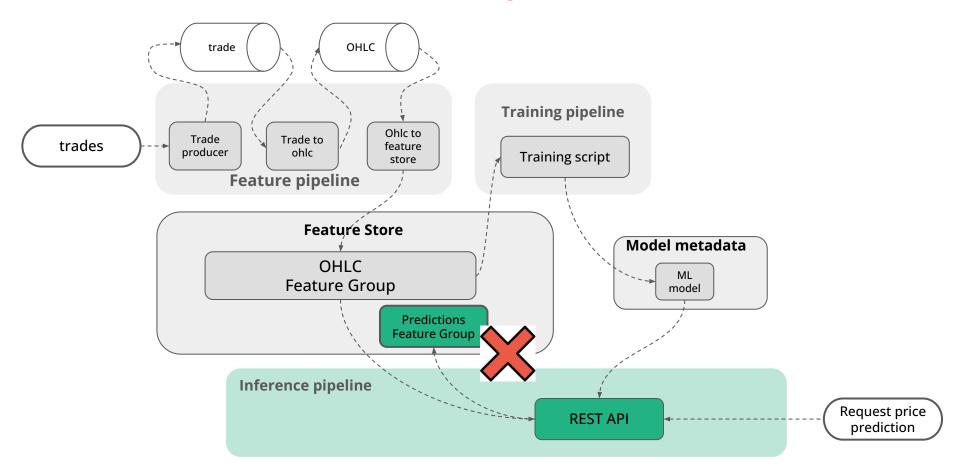
Fetch cached prediction



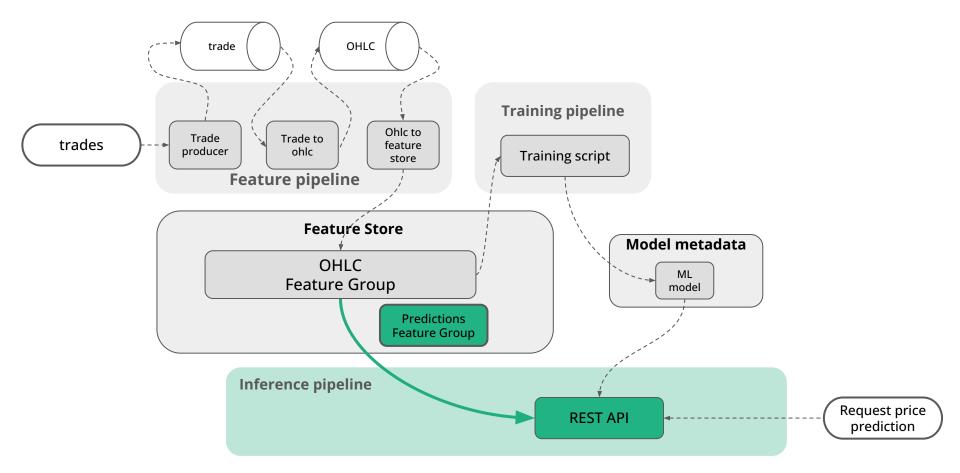
Return cached prediction

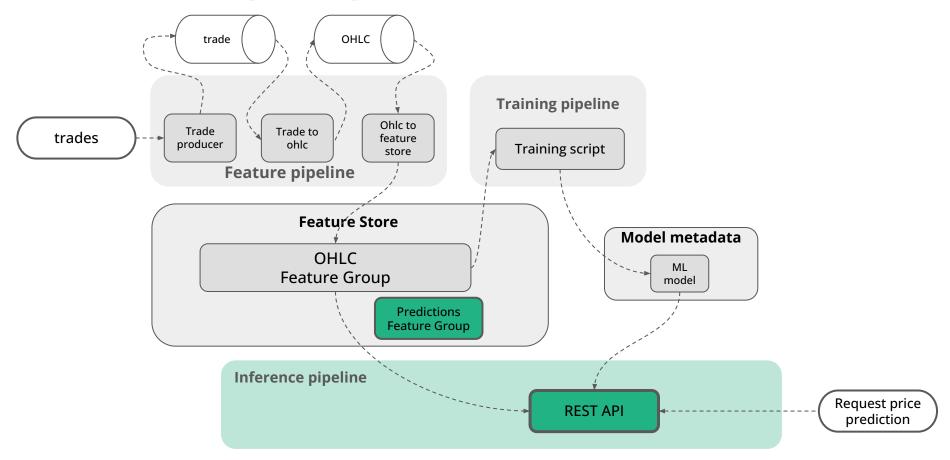


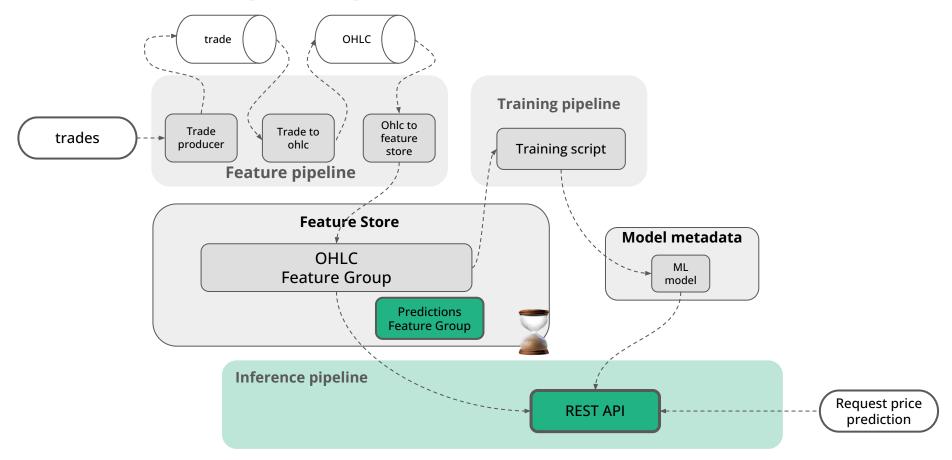
If NO

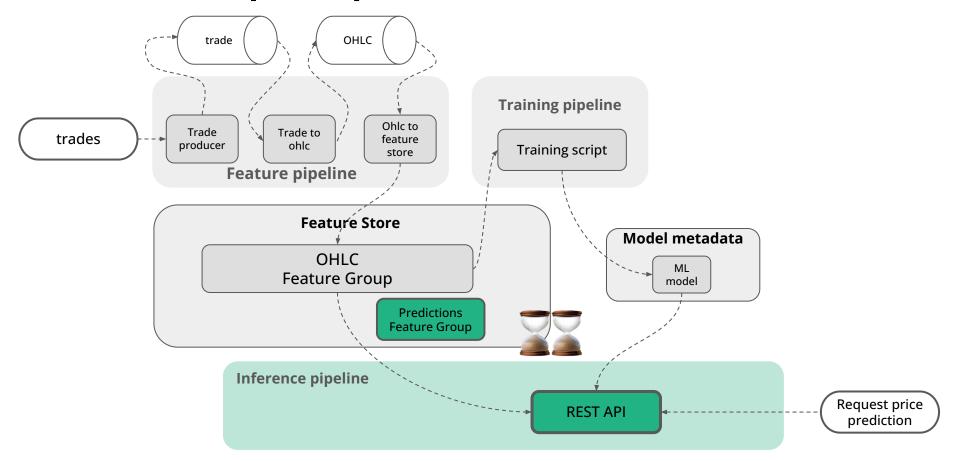


Gets fresh features from the store

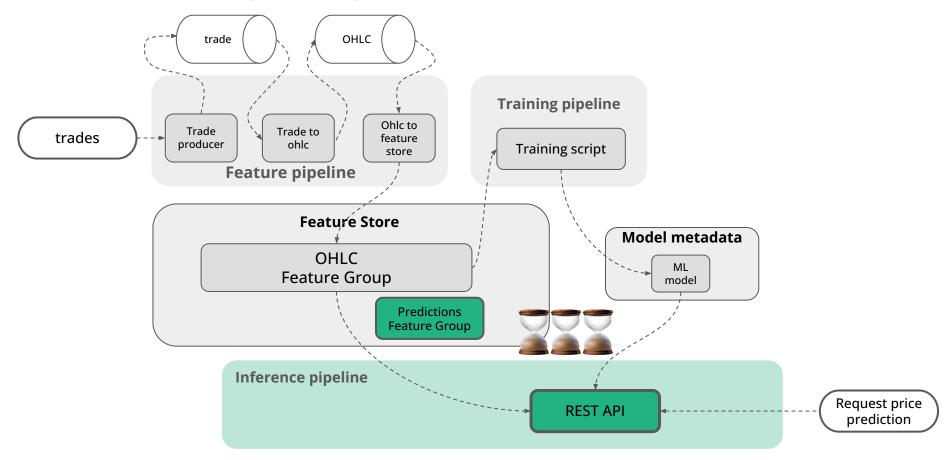




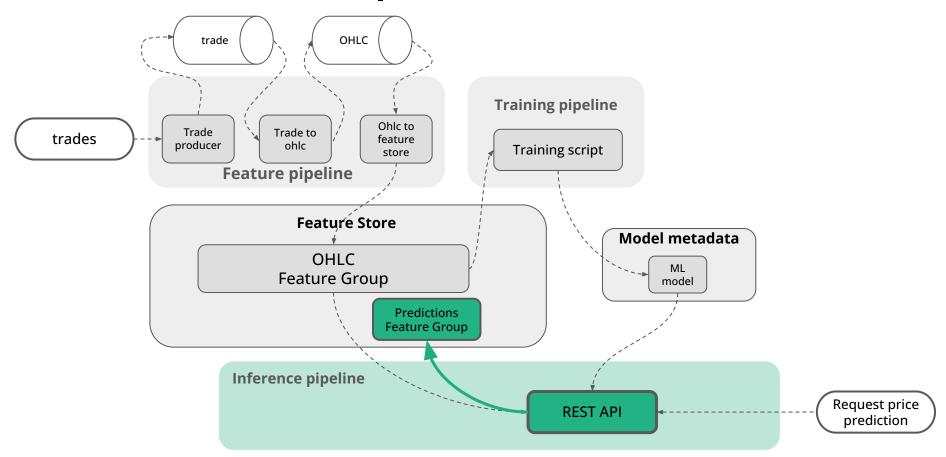




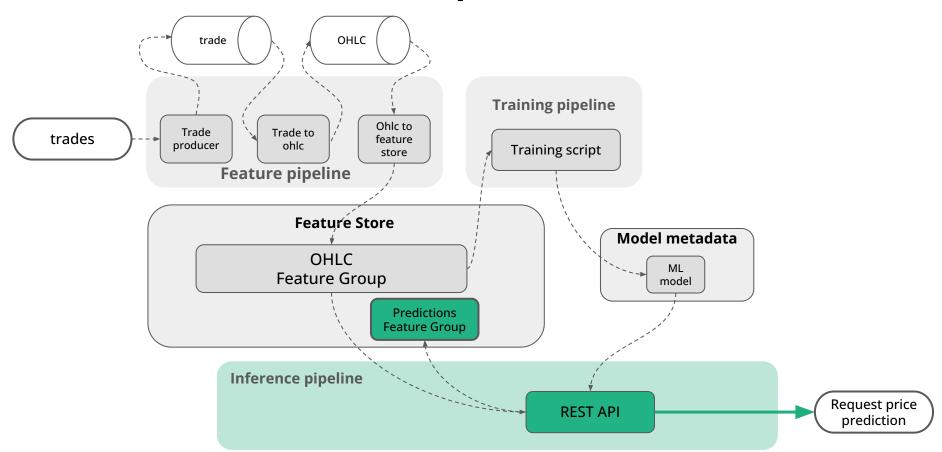
Computes prediction with ML Model



Pushes prediction to cache



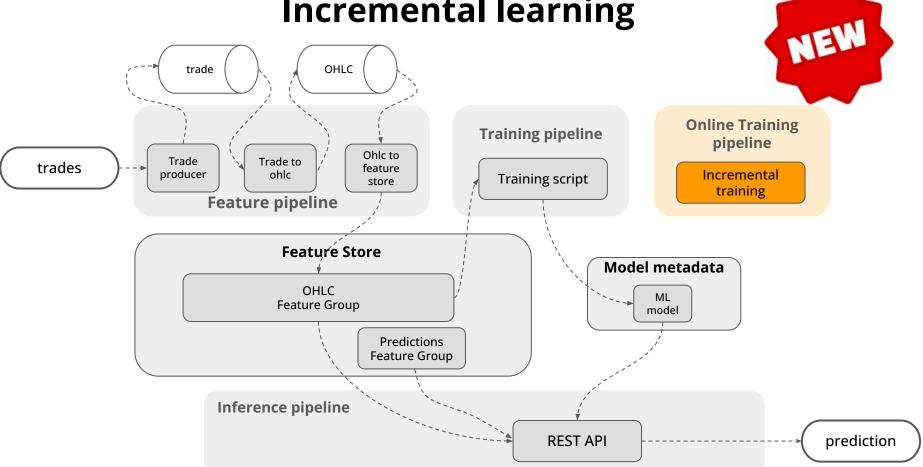
Returns prediction



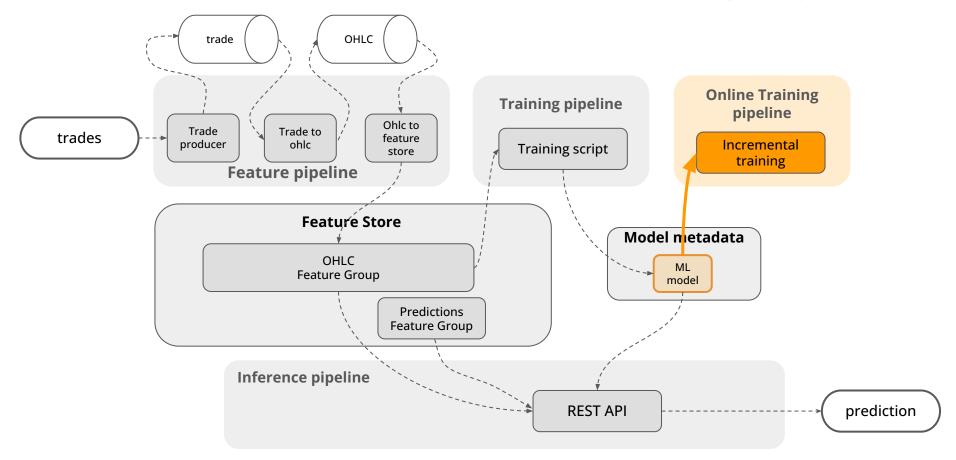


Incremental learning

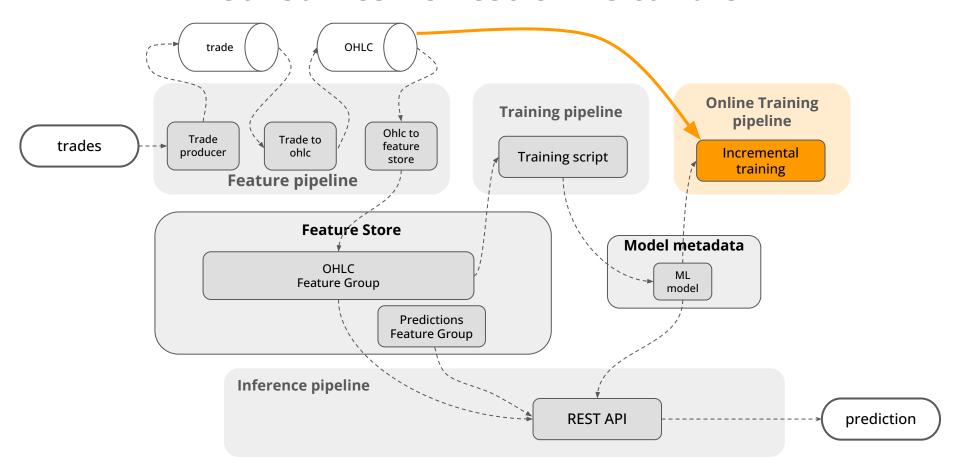
Incremental learning



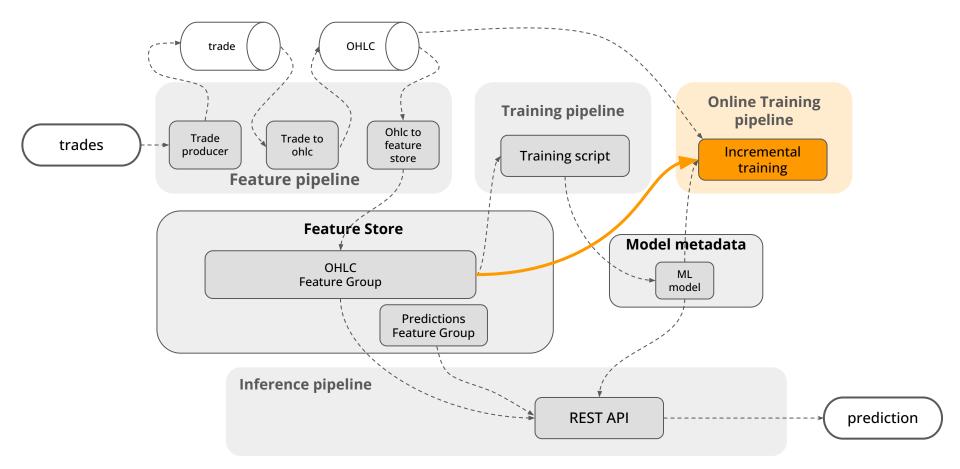
Loads PRODUCTION model from registry



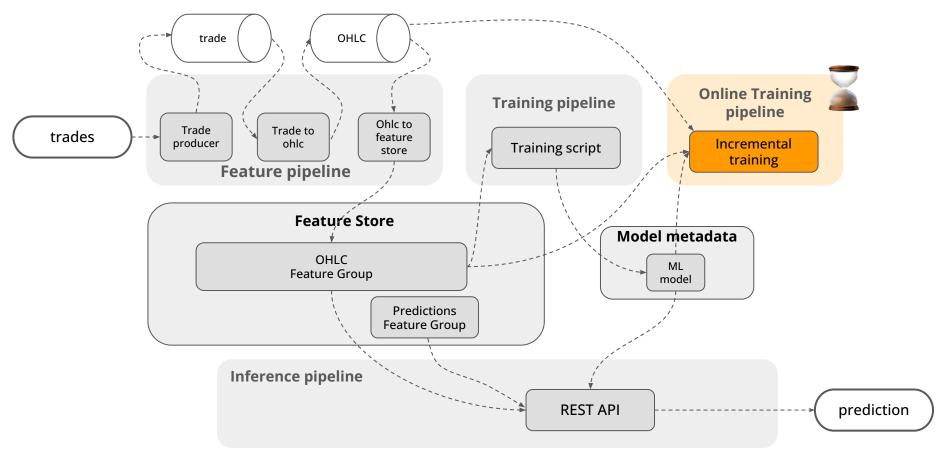
Consumes newest OHLC candle



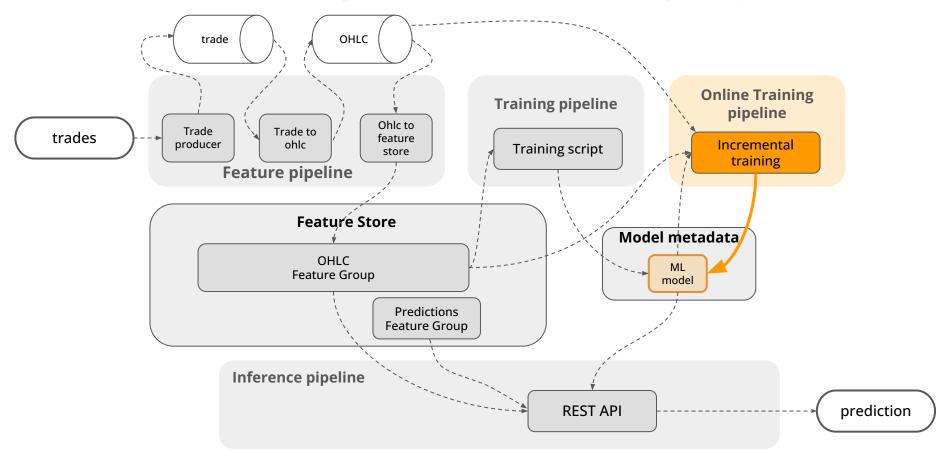
Fetches features from store



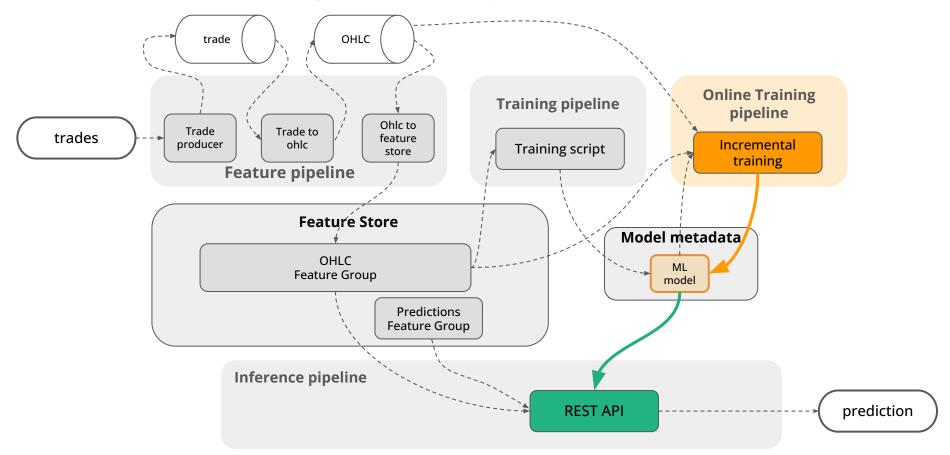
Updates the model parameters



Pushes updated model to registry



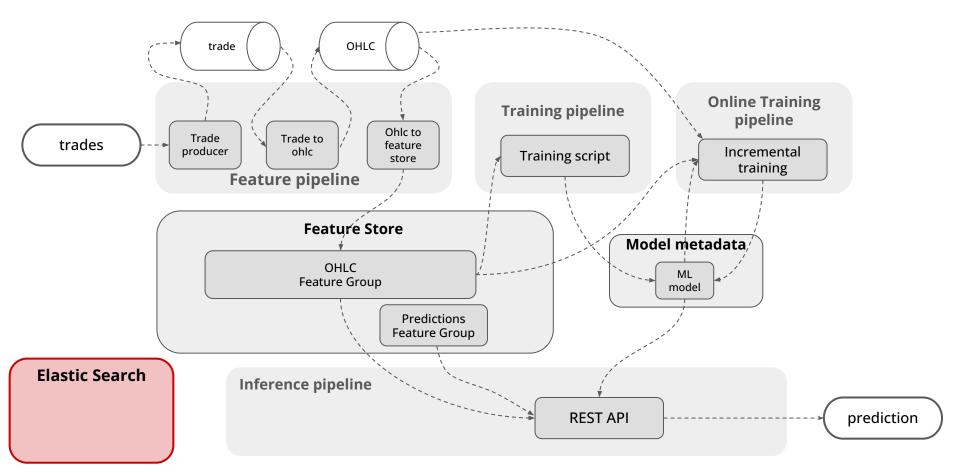
REST API periodically refreshes model



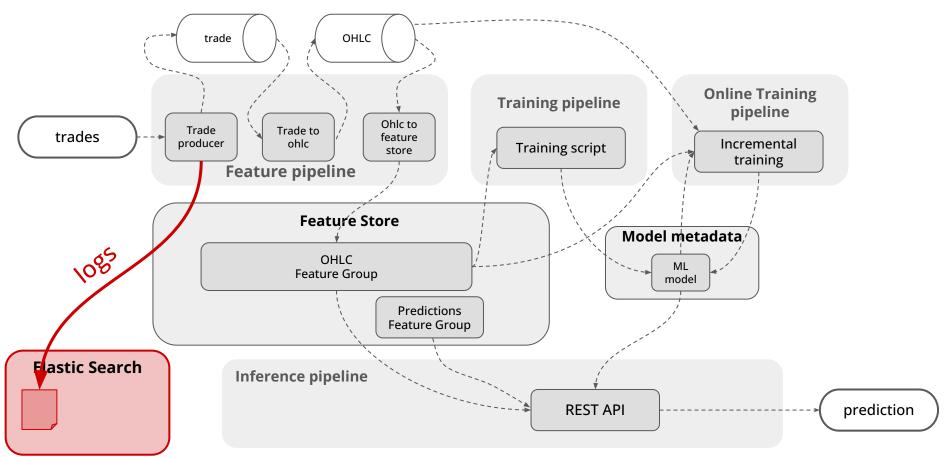


Monitoring

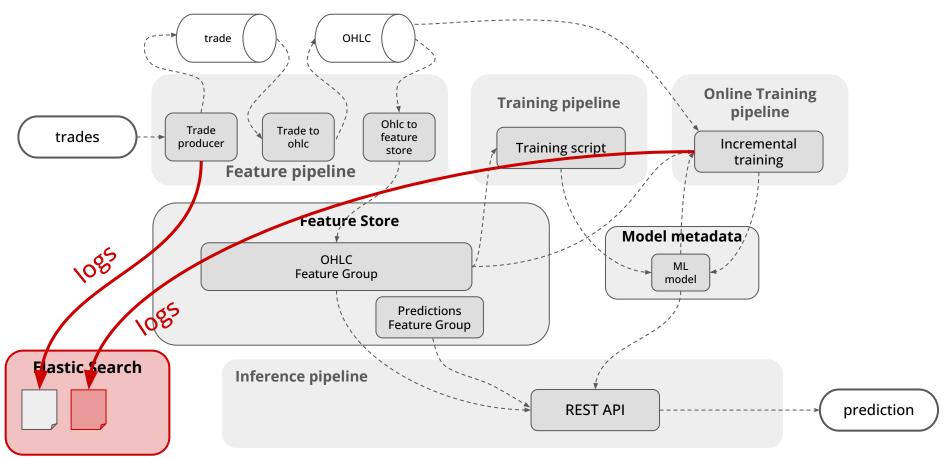
Add Elasticsearch cluster



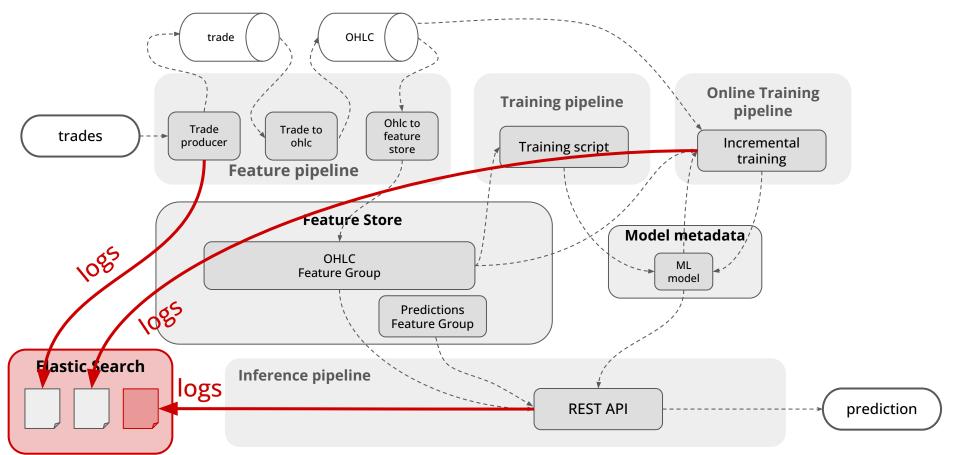
Logs for `trade_producer`



Logs for `incremental_trainer`

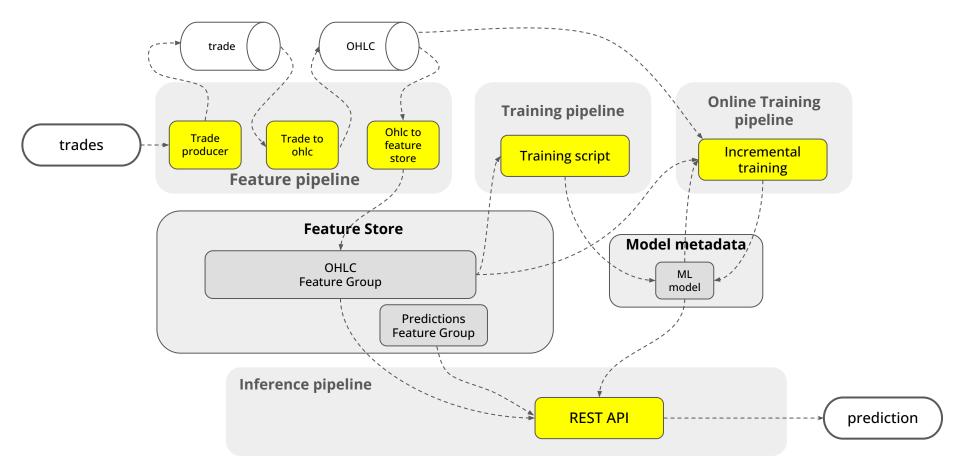


Logs for `REST API`

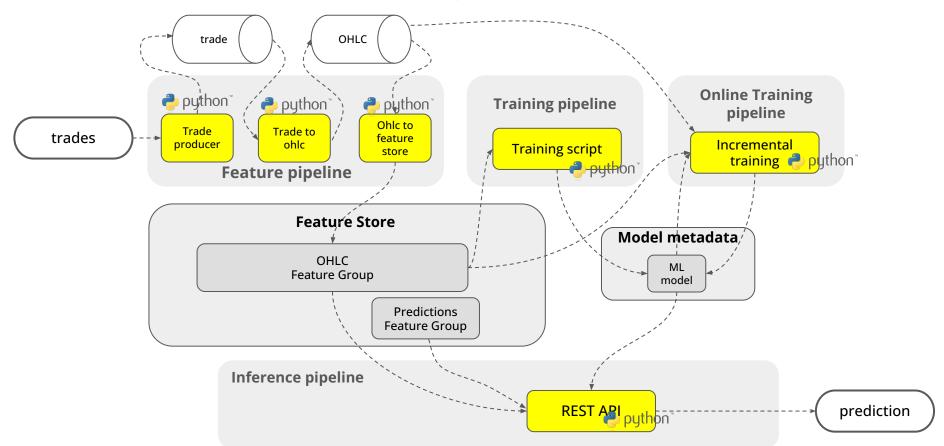


Production ready from DAY 1

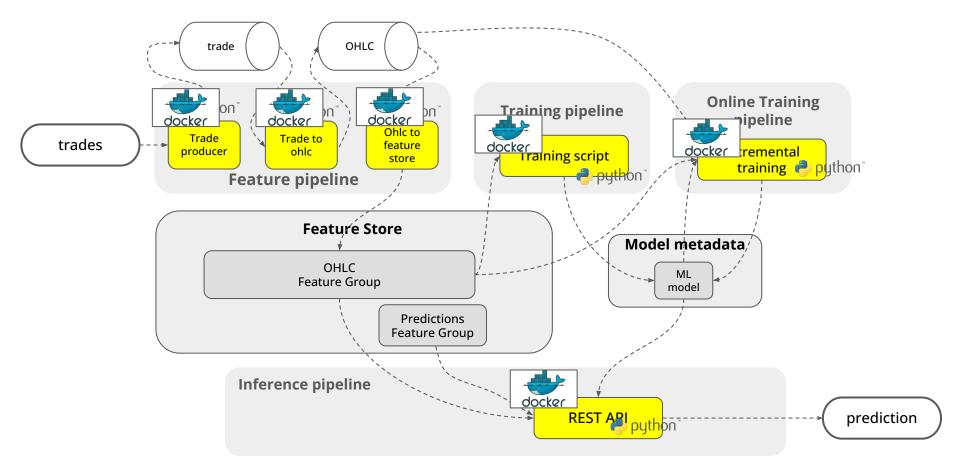
6 microservices



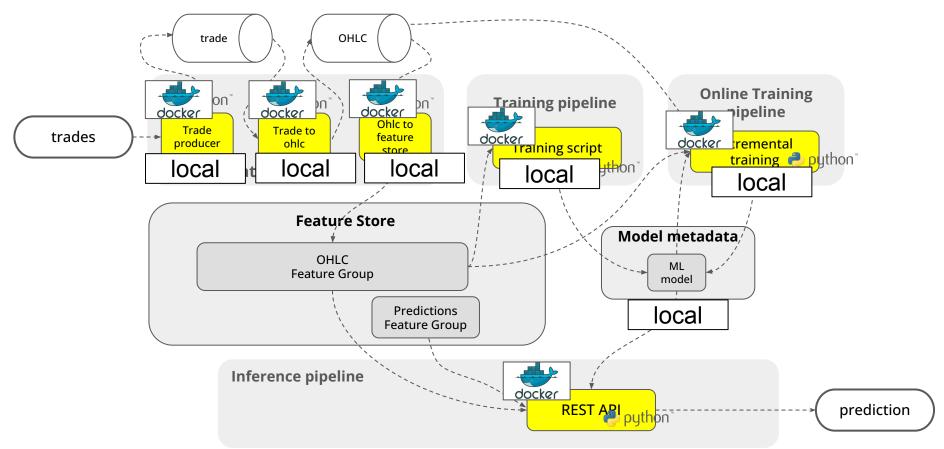
in Python



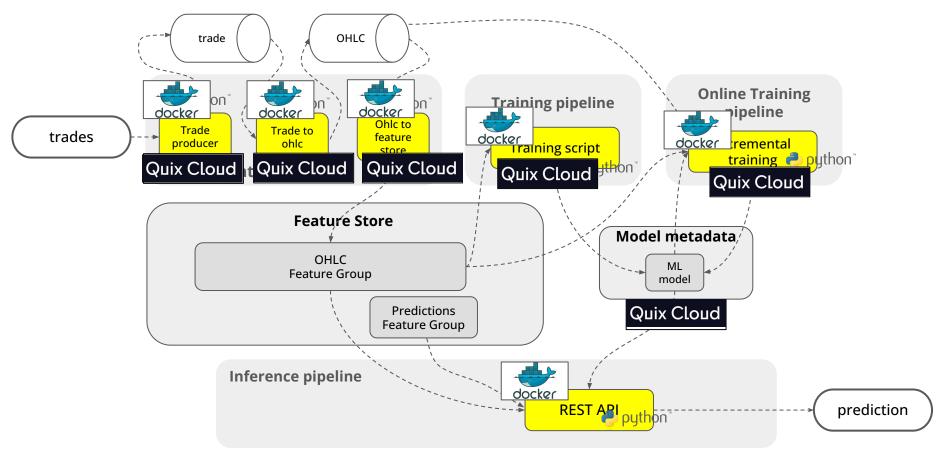
Dockerized



Develop locally



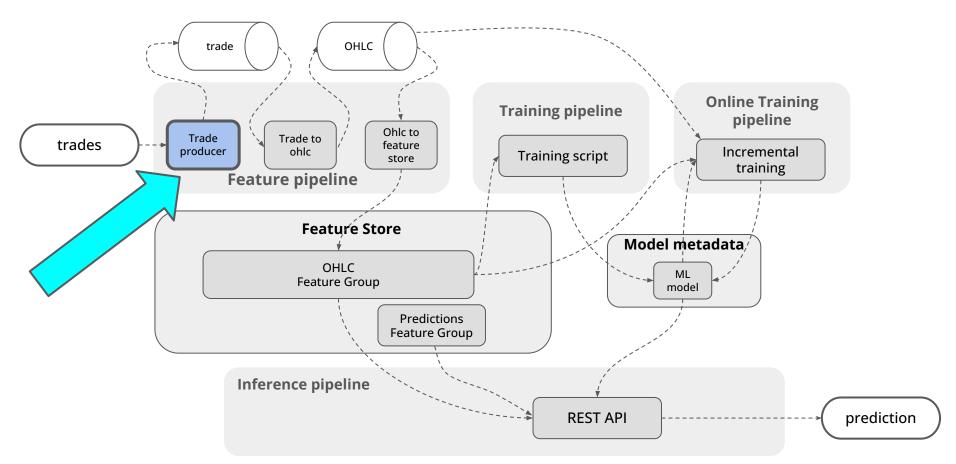
Deploy to Quix Cloud



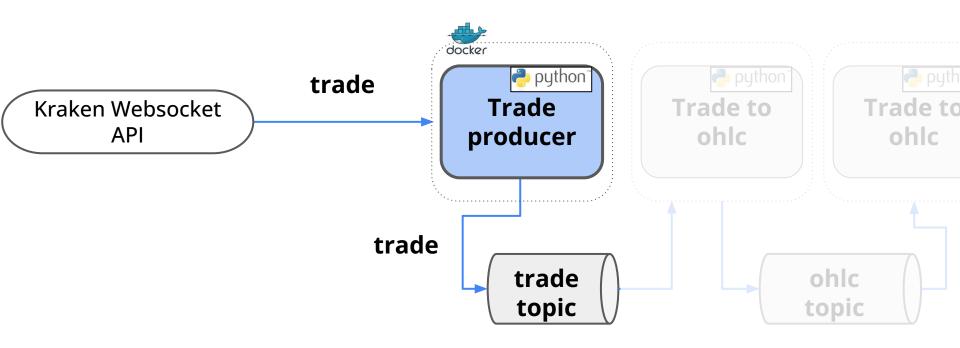
Let's start building!



This is where we start



Today we will build the trade producer



On Thursday we will build the other 2 steps

