Feature Store

Семинар 11

Определение и зачем надо

Feature Store – это централизованное хранилище для **хранения, управления и предоставления** признаков для моделей машинного обучения.

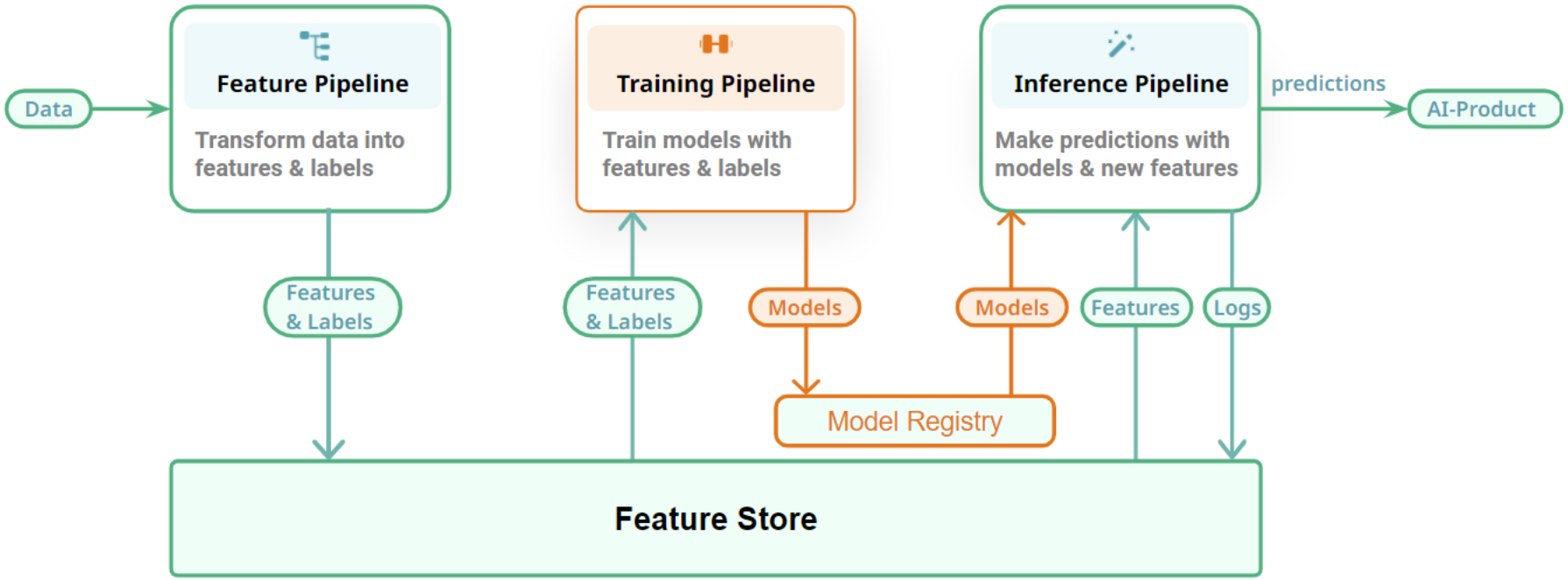
Нужен для:

1. Упрощения процесса извлечения и подготовки данных
2. Повторного использования признаков между проектами
3. Обеспечения согласованности и версионности данных

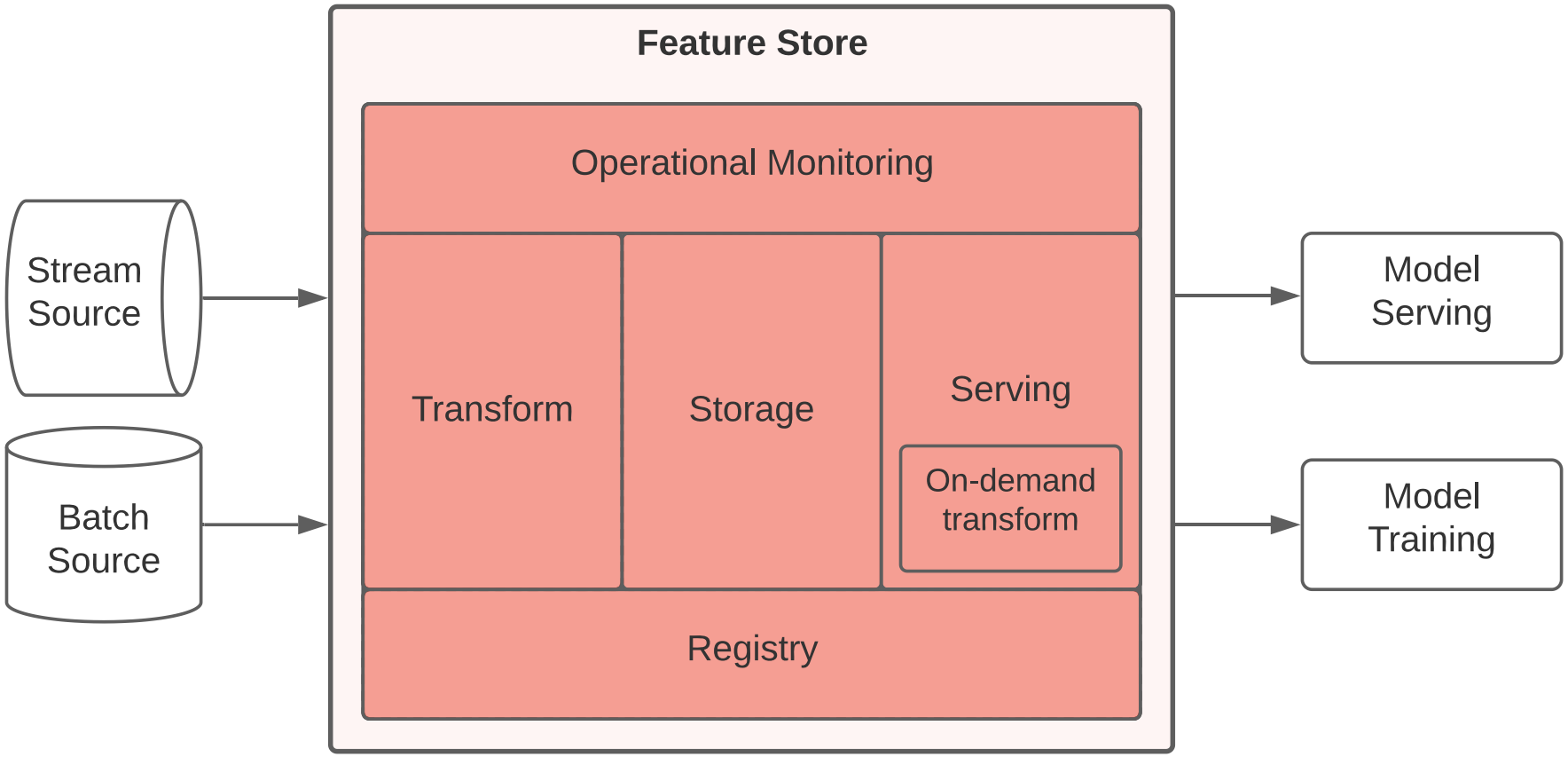
Где можно почитать

- https://www.featurestore.org
- https://www.hopsworks.ai/dictionary/feature-store
- https://habr.com/ru/companies/glowbyte/articles/581458/
- https://feast.dev/blog/what-is-a-feature-store/

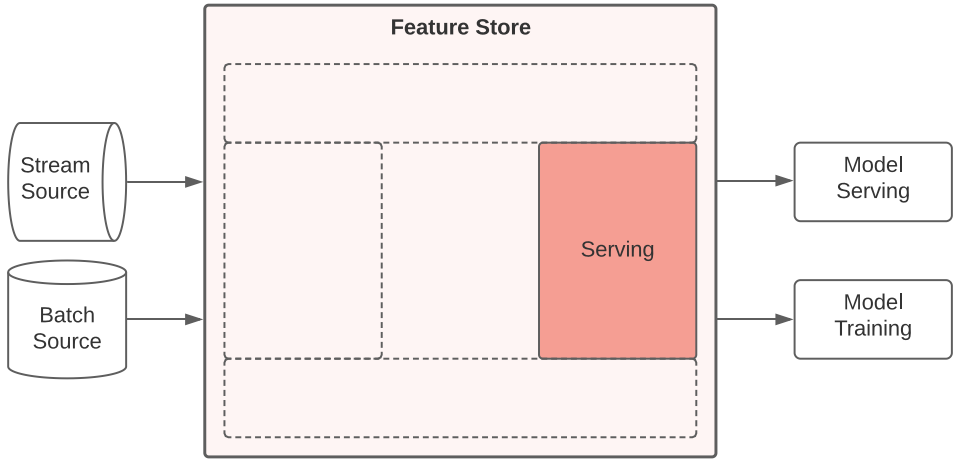
Как выглядит в процессах

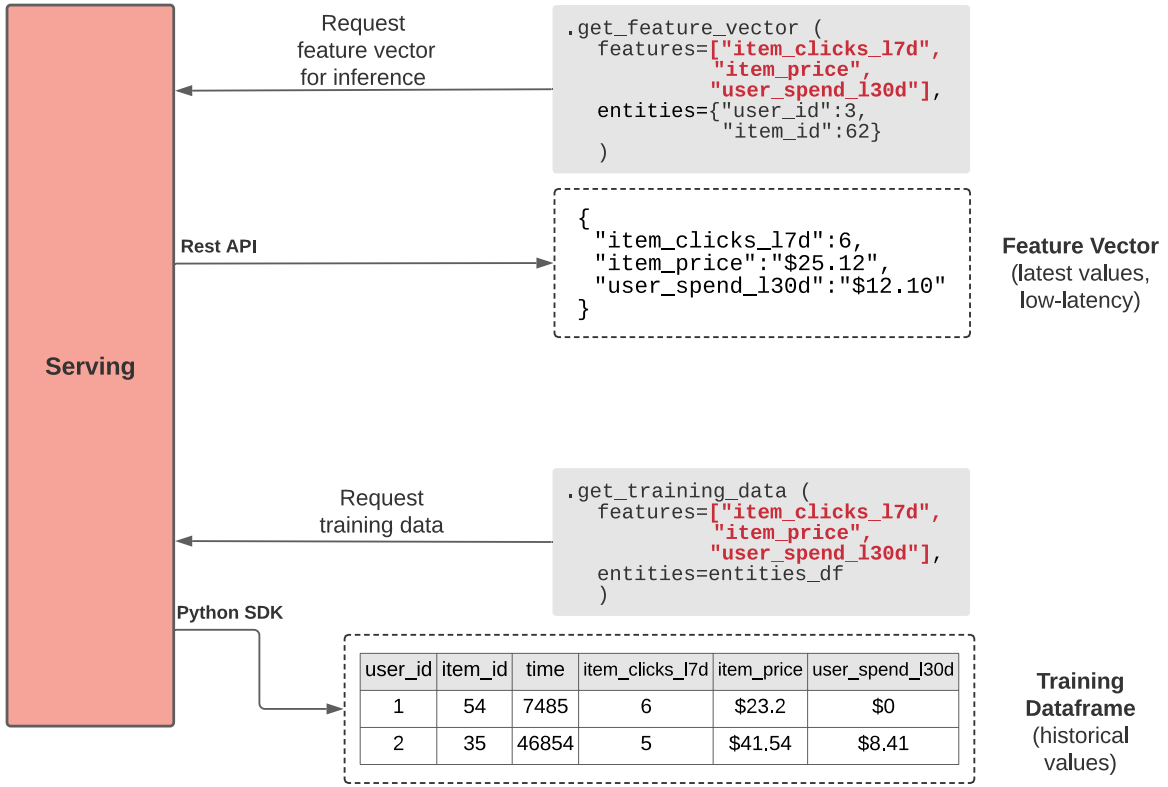


Компоненты FS

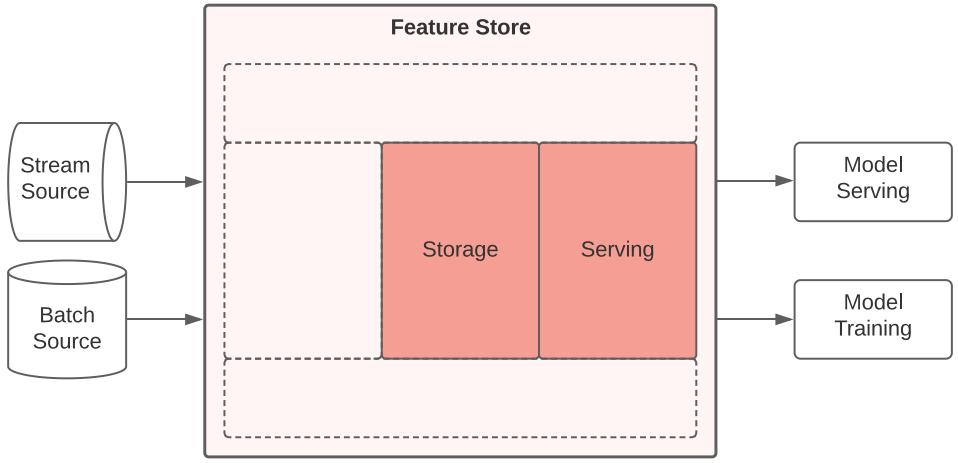


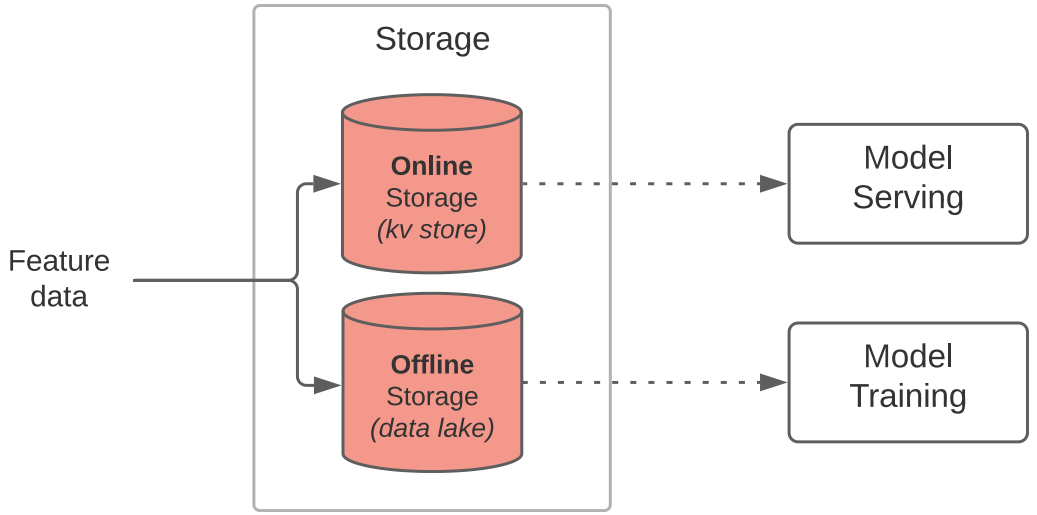
1. Serving


Feature Store
Stream Source
Batch Source
Serving
Model Serving
Model Training

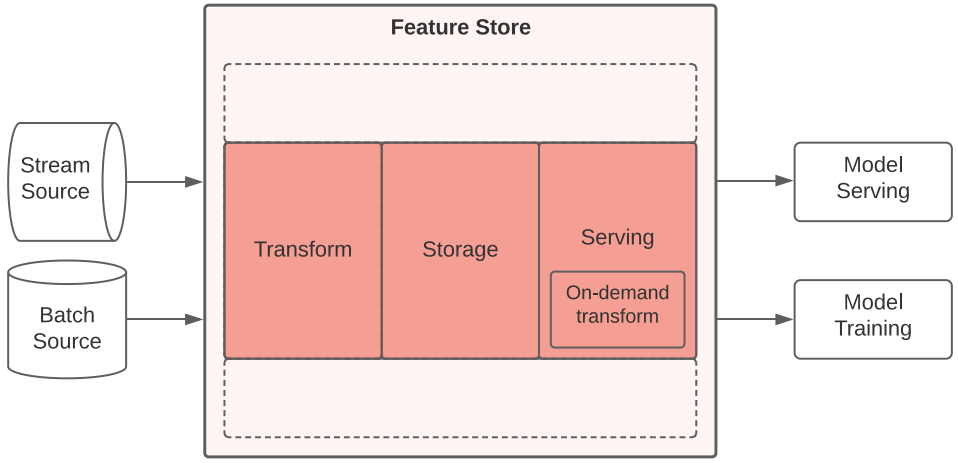

Serving
Request feature vector for inference
.get_feature_vector (
features=["item_clicks_l7d", "item_price", "user_spend_l30d"],
entities={"user_id":3, "item_id":62}
)
Rest API
{
"item_clicks_l7d":6,
"item_price":"$25.12",
"user_spend_l30d":"$12.10"
}
Feature Vector (latest values, low-latency)
Request training data
.get_training_data (
features=["item_clicks_l7d", "item_price", "user_spend_l30d"],
entities=entities_df
)
Python SDK
user_id item_id time item_clicks_l7d item_price user_spend_l30d
1 54 7485 6 $23.2 $0
2 35 46854 5 $41.54 $8.41
Training Dataframe (historical values)

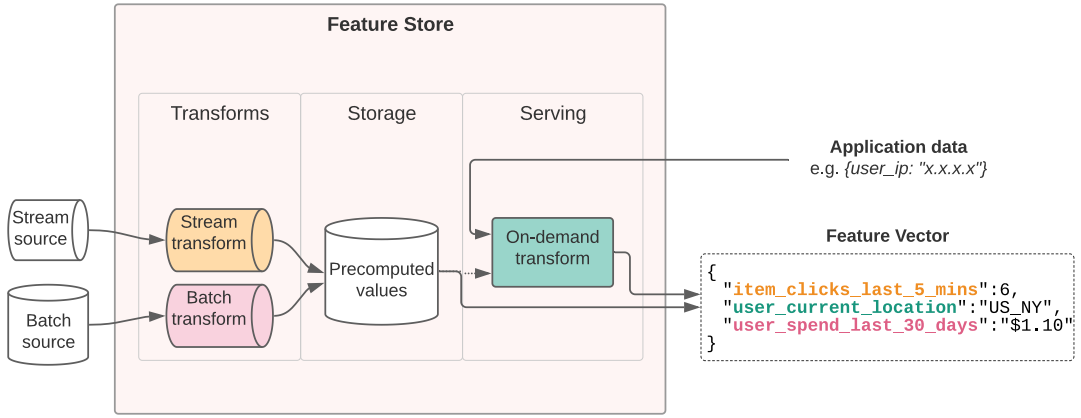
2. Storage


Feature Store
Stream Source
Batch Source
Storage
Serving
Model Serving
Model Training

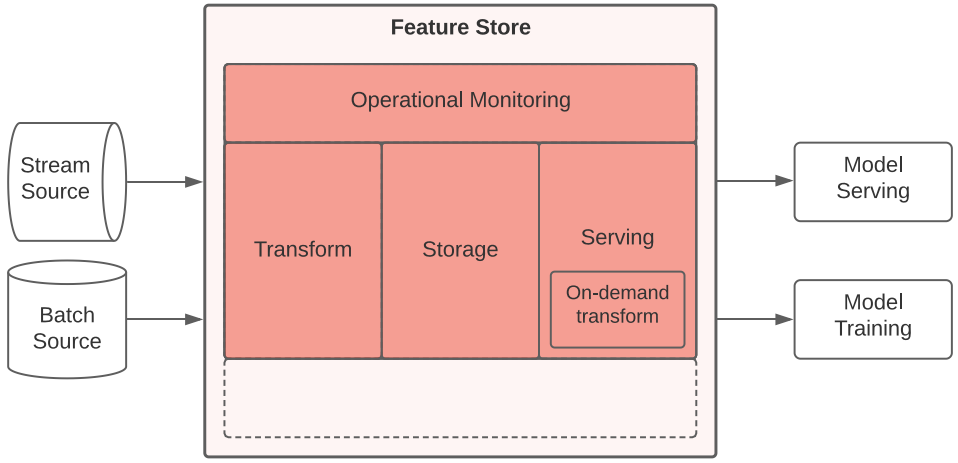

Storage
Feature data
Online Storage (kv store)
Offline Storage (data lake)
Model Serving
Model Training

3. Transformation





4. Monitoring



Dataset profile

Currently, Feast supports only Great Expectation's ExpectationSuite as dataset's profile. Hence, the user needs to define a function (profiler) that would receive a dataset and return an ExpectationSuite.

Great Expectations supports automatic profiling as well as manually specifying expectations:

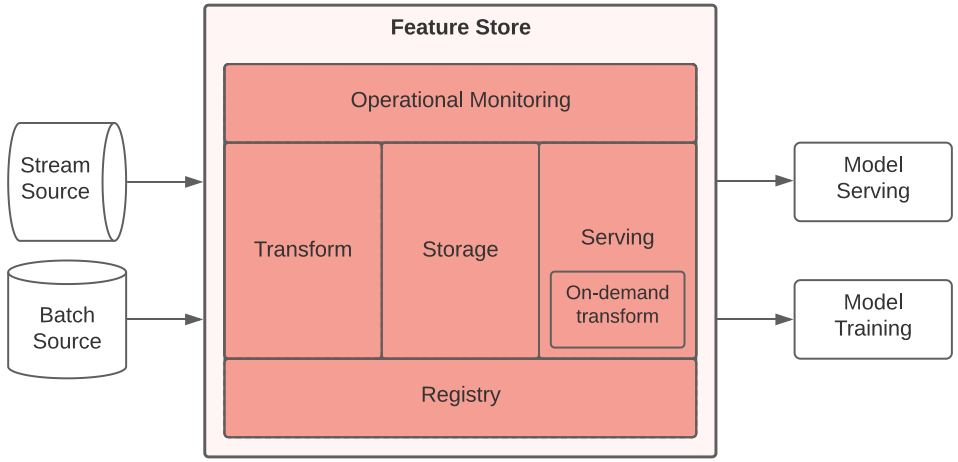
```
from great_expectations.dataset import Dataset
from great_expectations.core.expectation_suite import ExpectationSuite

from feast.dqm.profilers.ge_profiler import ge_profiler

@ge_profiler
def automatic_profiler(dataset: Dataset) -> ExpectationSuite:
    from great_expectations.profile.user_configurable_profiler import UserConfigurablePro

    return UserConfigurableProfiler(
        profile_dataset=dataset,
        ignored_columns=['conv_rate'],
        value_set_threshold='few'
    ).build_suite()
```

5. Registry

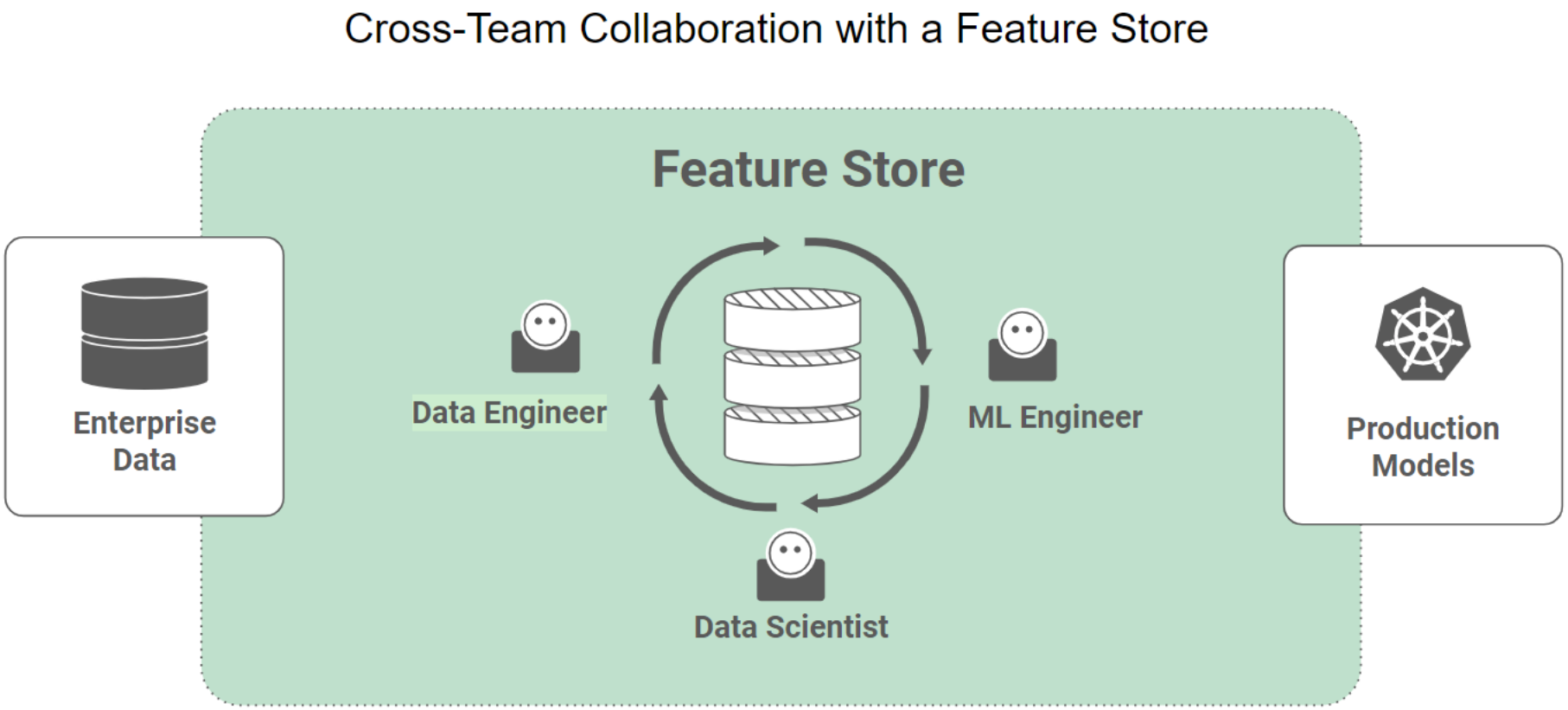


Registries

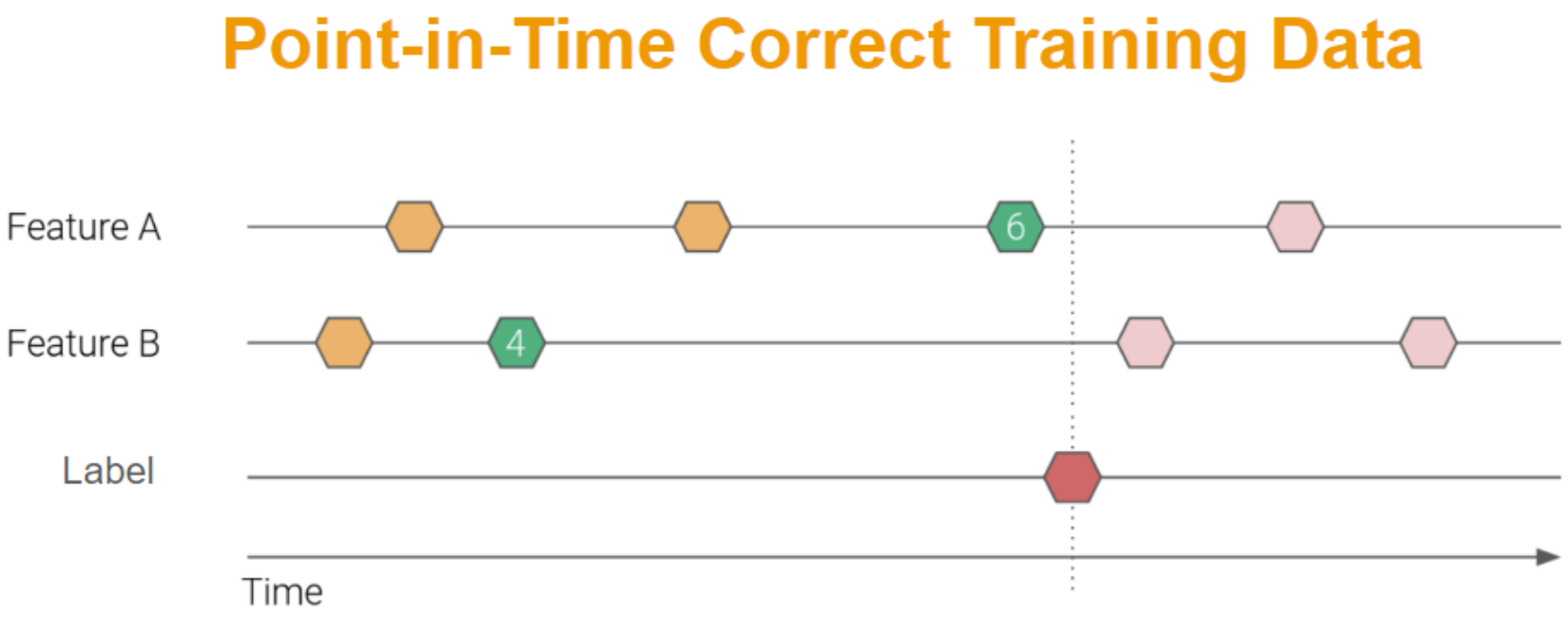
Please see Registry for a conceptual explanation of registries.

- Local
- S3
- GCS
- SQL
- Snowflake

Централизованный доступ



Сбор признаков по таргету



Смотрим Feast

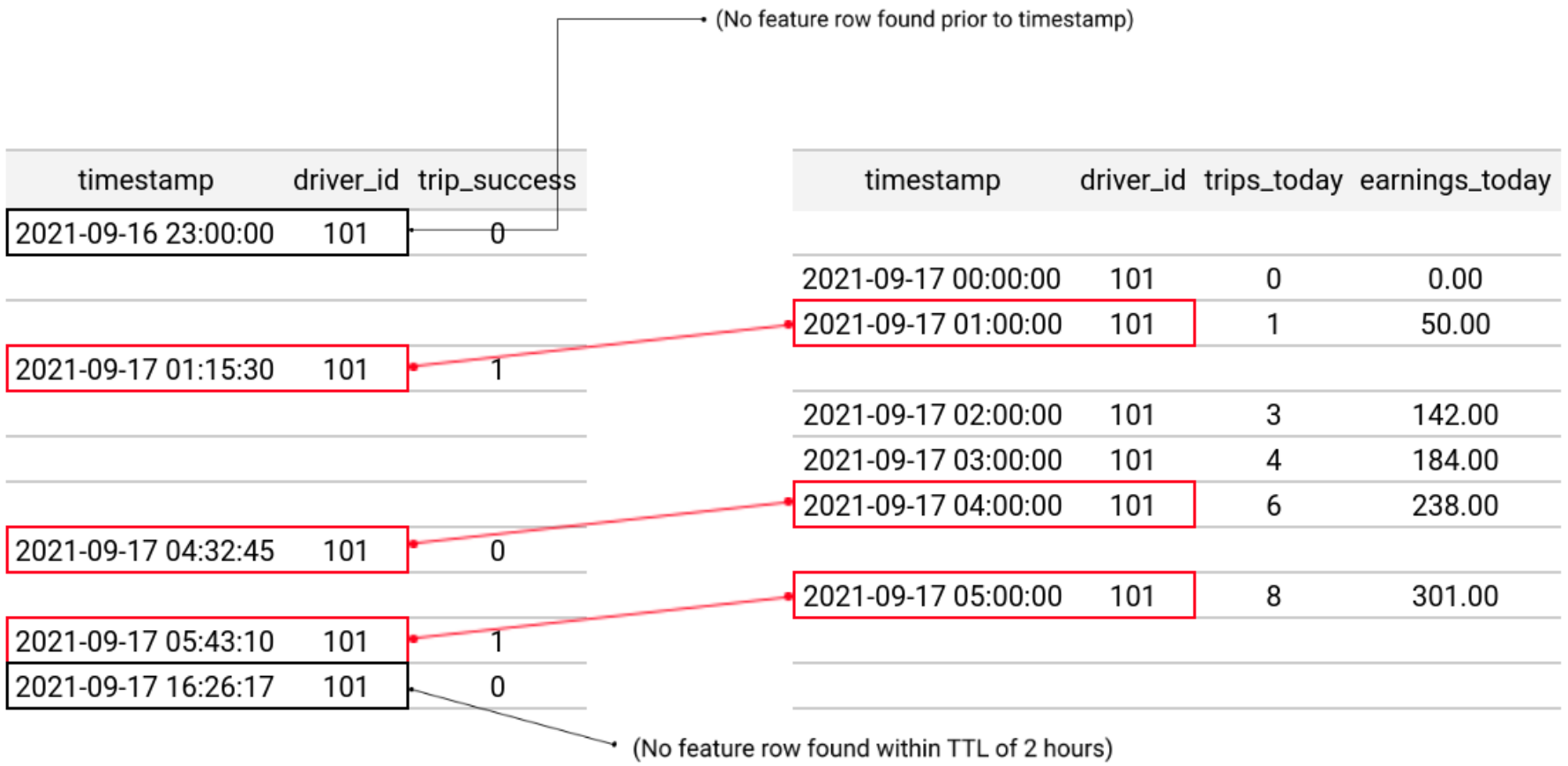
https://docs.feast.dev/getting-started/quickstart

Point-in-time joins

timestamp	driver_id	trips_today	earnings_today
2021-09-17 00:00:00	101	0	0.00
2021-09-17 01:00:00	101	1	50.00
2021-09-17 02:00:00	101	3	142.00
2021-09-17 03:00:00	101	4	184.00
2021-09-17 04:00:00	101	6	238.00
2021-09-17 05:00:00	101	8	301.00

timestamp	driver_id	trip_success
2021-09-16 23:00:00	101	0
2021-09-17 01:15:30	101	1
2021-09-17 04:32:45	101	0
2021-09-17 05:43:10	101	1
2021-09-17 16:26:17	101	0

Point-in-time joins



Point-in-time joins

timestamp	driver_id	trip_success	trips_today	earnings_today
2021-09-16 23:00:00	101	0	NULL	NULL
2021-09-17 01:15:30	101	1	1	50.00
2021-09-17 04:32:45	101	0	6	238.00
2021-09-17 05:43:10	101	1	8	301.00
2021-09-17 16:26:17	101	0	NULL	NULL

Домашняя работа №6 / Финальный проект

1 балл:

1. Поднять локально Feast (как в примере сегодня)
2. Собрать данные за 10 таймстемпов (любых, для всех driver_id) и обучить модель предсказывать колонку **avg_daily_trips** по полям **conv_rate** и **acc_rate**

2 балла:

1. Поднять локально Feast + Airflow
2. Написать DAG для подтягивания актуальных фичей за 10 таймстемпов (любых) для всех driver_id и обучение модели (как в задаче за 1 балл)

Домашняя работа №6 / Финальный проект

3 балла:

1. Все то, что было в задании на 2 балла.
2. Написать DAG для батч инференса модели (должен уметь делать инференс как на актуальную дату, так и на любую дату в прошлом). Модель сохранять в MLflow

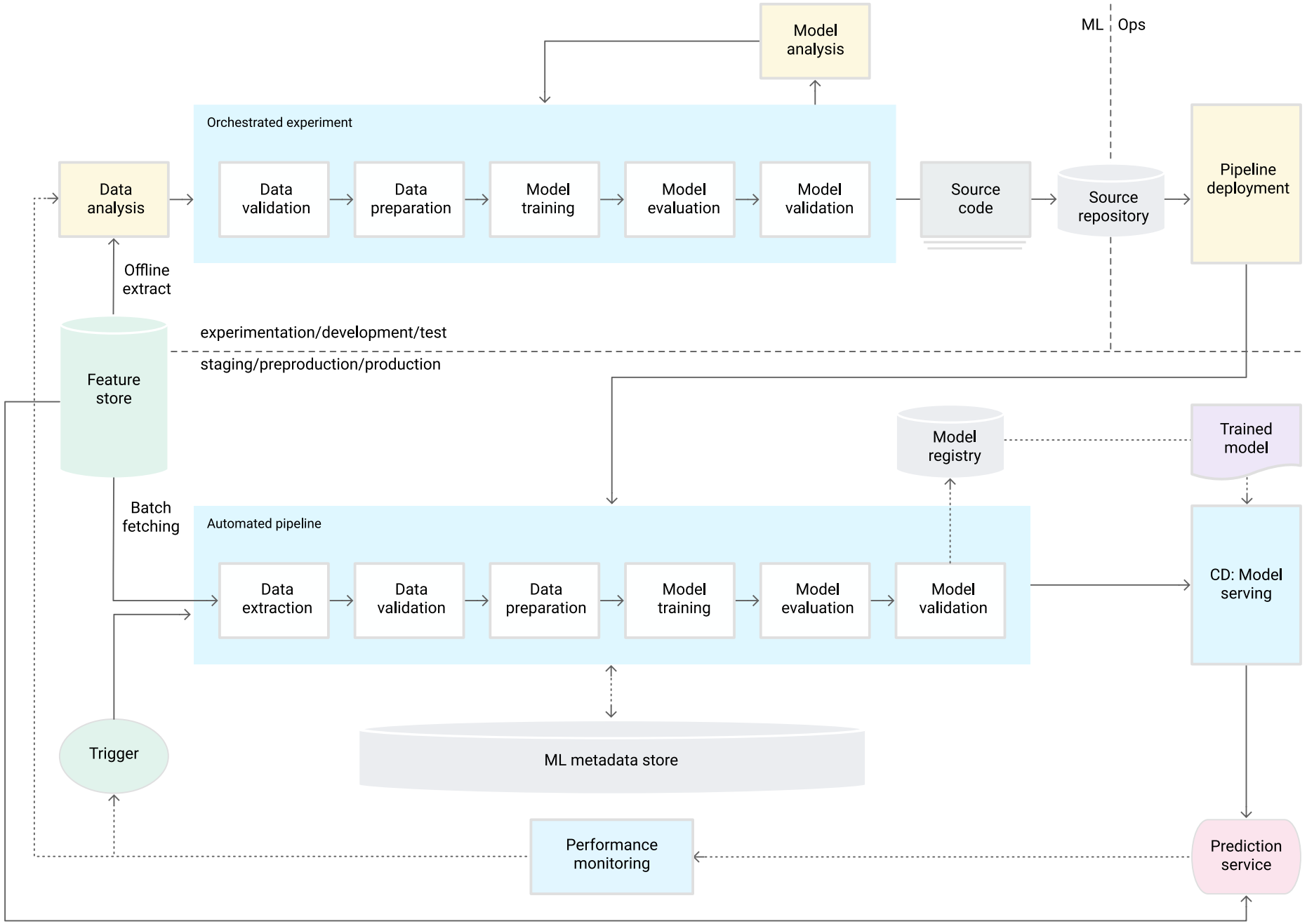
4 балла:

1. Все то, что было в задании на 3 балла.
2. Поднять сервис с онлайн инференсом модели
3. Забирать онлайн фичи из FS и отдавать на вход сервису с онлайн инференсом
4. * (опционально) прикрутить мониторинг входных и выходных данных

Домашняя работа №6 / Финальный проект

https://github.com/feast-dev/feast-workshop/tree/main/module_1

Для вдохновения


Model analysis
ML
Ops
Orchestrated experiment
Data analysis
Data validation
Data preparation
Model training
Model evaluation
Model validation
Source code
Source repository
Pipeline deployment
Offline extract
experimentation/development/test
staging/preproduction/production
Feature store
Model registry
Trained model
Batch fetching
Automated pipeline
Data extraction
Data validation
Data preparation
Model training
Model evaluation
Model validation
CD: Model serving
Trigger
ML metadata store
Performance monitoring
Prediction service

Итоговые баллы

В итоге 14 баллов за семестр

Для автомата надо сдать **все** лабы хотя бы на один балл

«5» = 11+ баллов

«4» = 8+ баллов

«3» = 6+ баллов