

Промышленный Мониторинг качества данных в Feature Store

Предпосылки и реализация



Алексей Лямзин

Дата-аналитик

Группа финтех аналитики

СОДЕРЖАНИЕ

- 01** Предпосылки
- 02** Существующие решения
- 03** Реализация
- 04** Примеры
- 05** Выводы
- 06** Q&A

| 1

Предпосылки

Предпосылки

Растет:



число сотрудников, работающих с данными

Предпосылки

Растет:



число сотрудников, работающих с данными



число экспериментов



число моделей и признаков

Растет:



число сотрудников, работающих с данными



число экспериментов

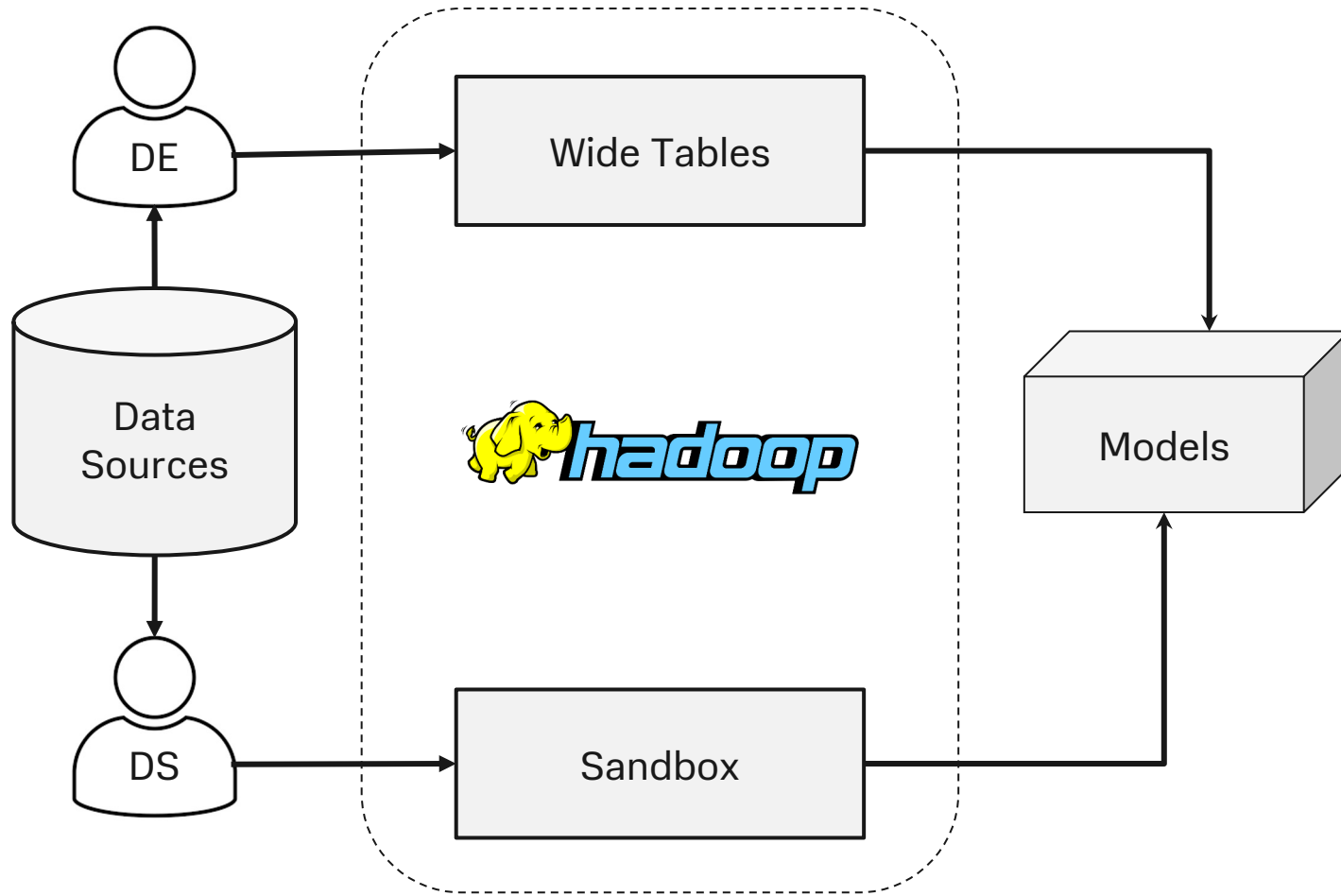


число моделей и признаков

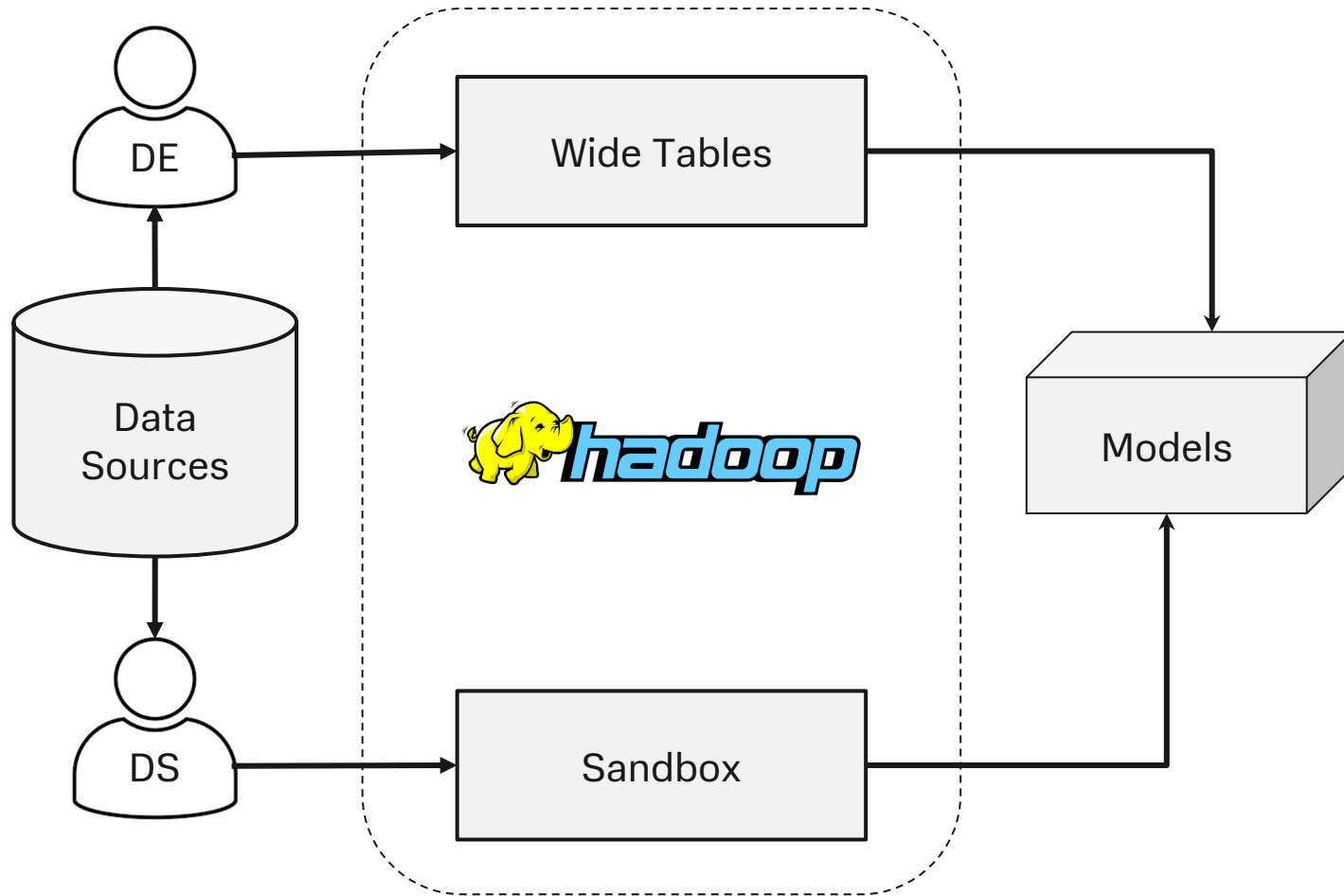


Масштабирование дается все
сложнее

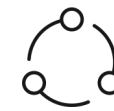
Предпосылки



Предпосылки



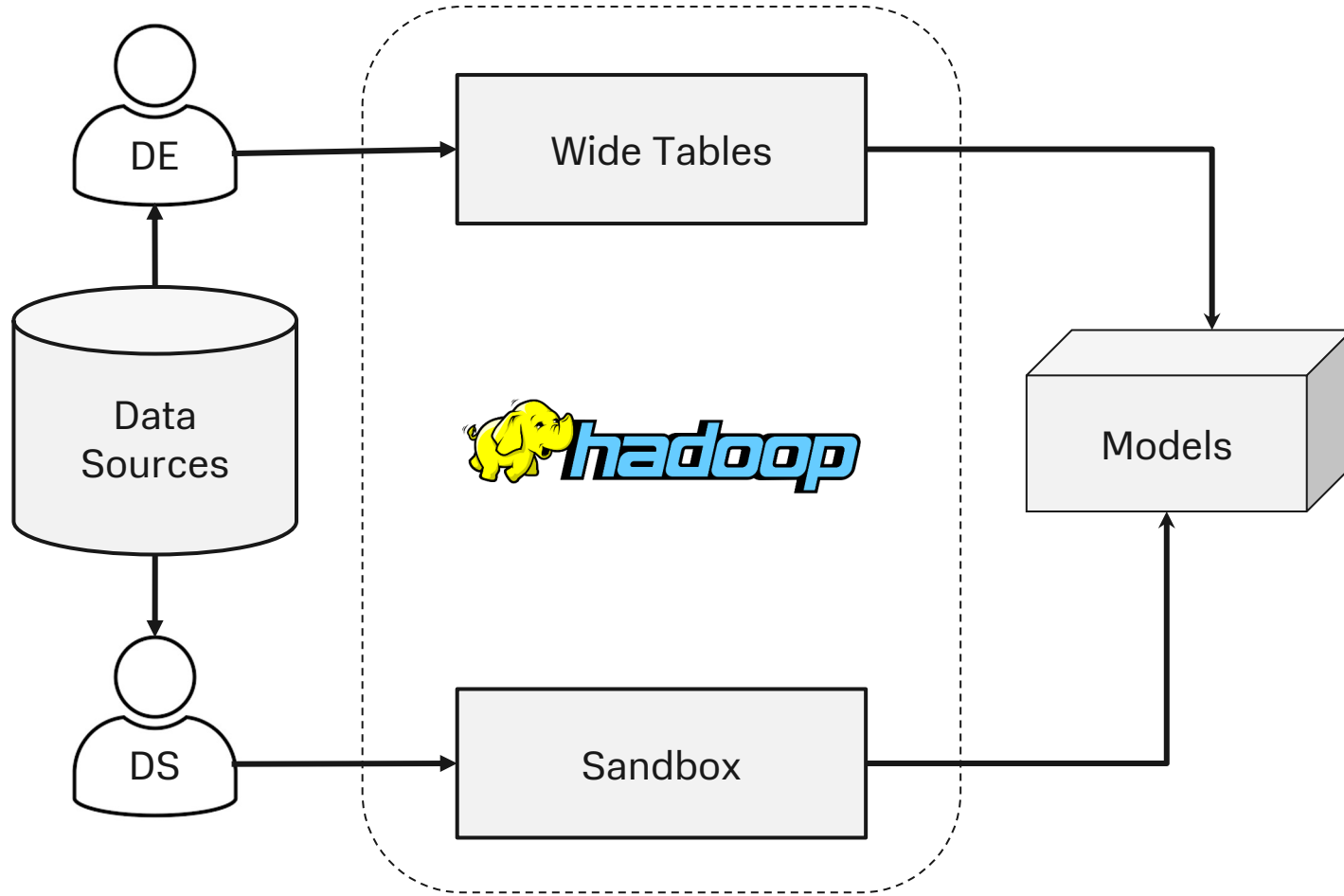
Проблемы:



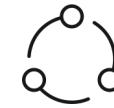
Команды делают одно и то же



Предпосылки



Проблемы:

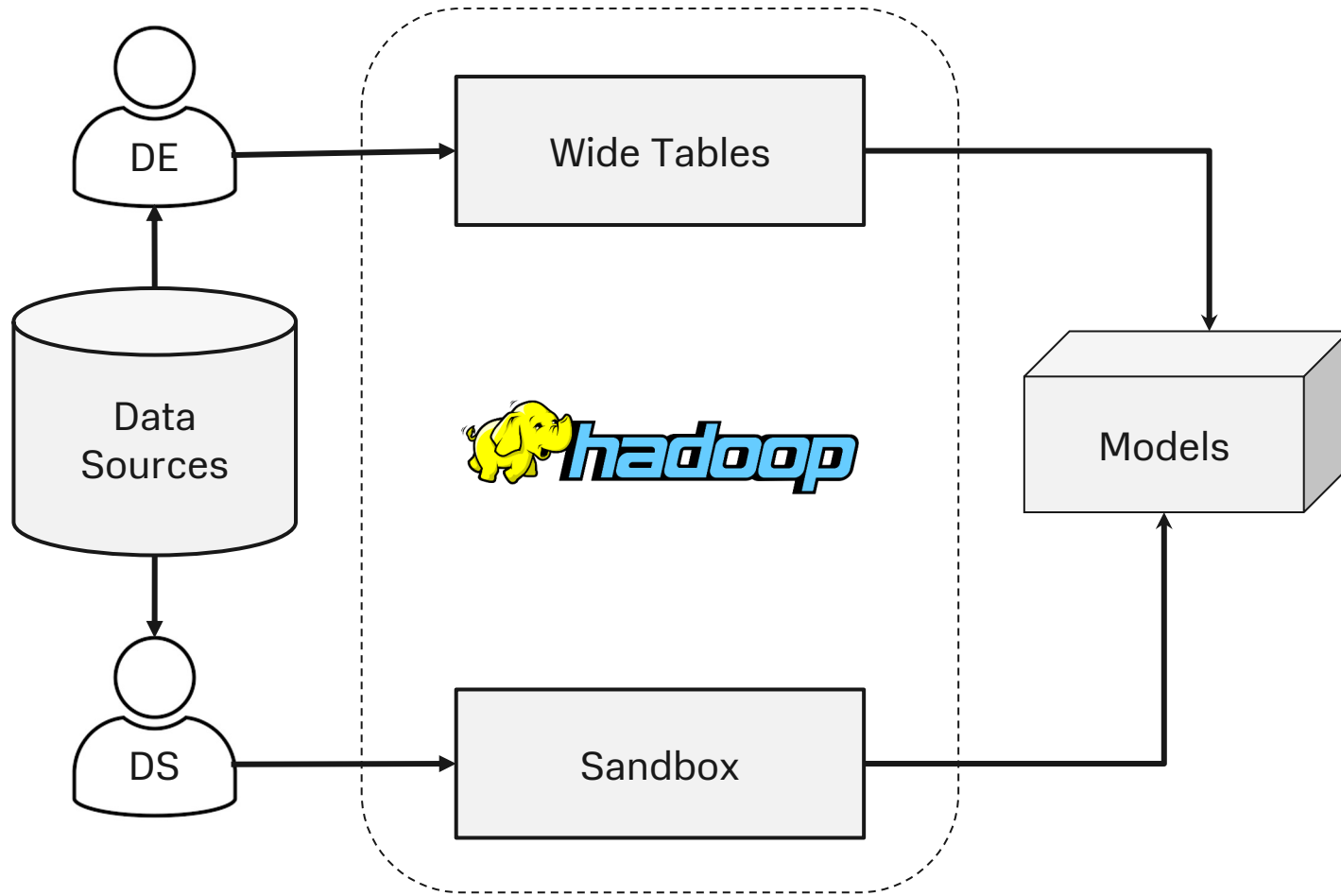


Команды делают одно и то же






Долгий доступ

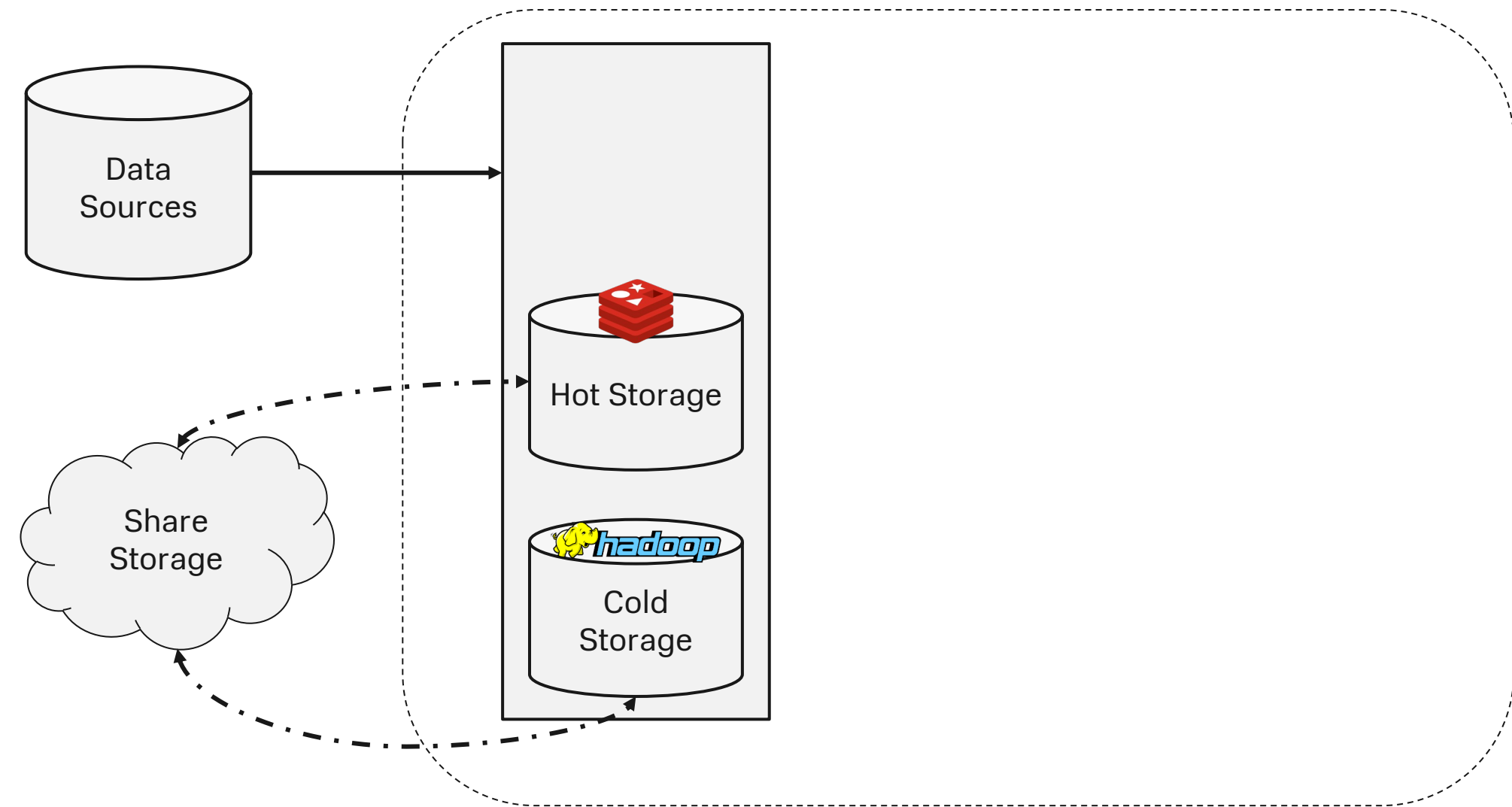
Предпосылки



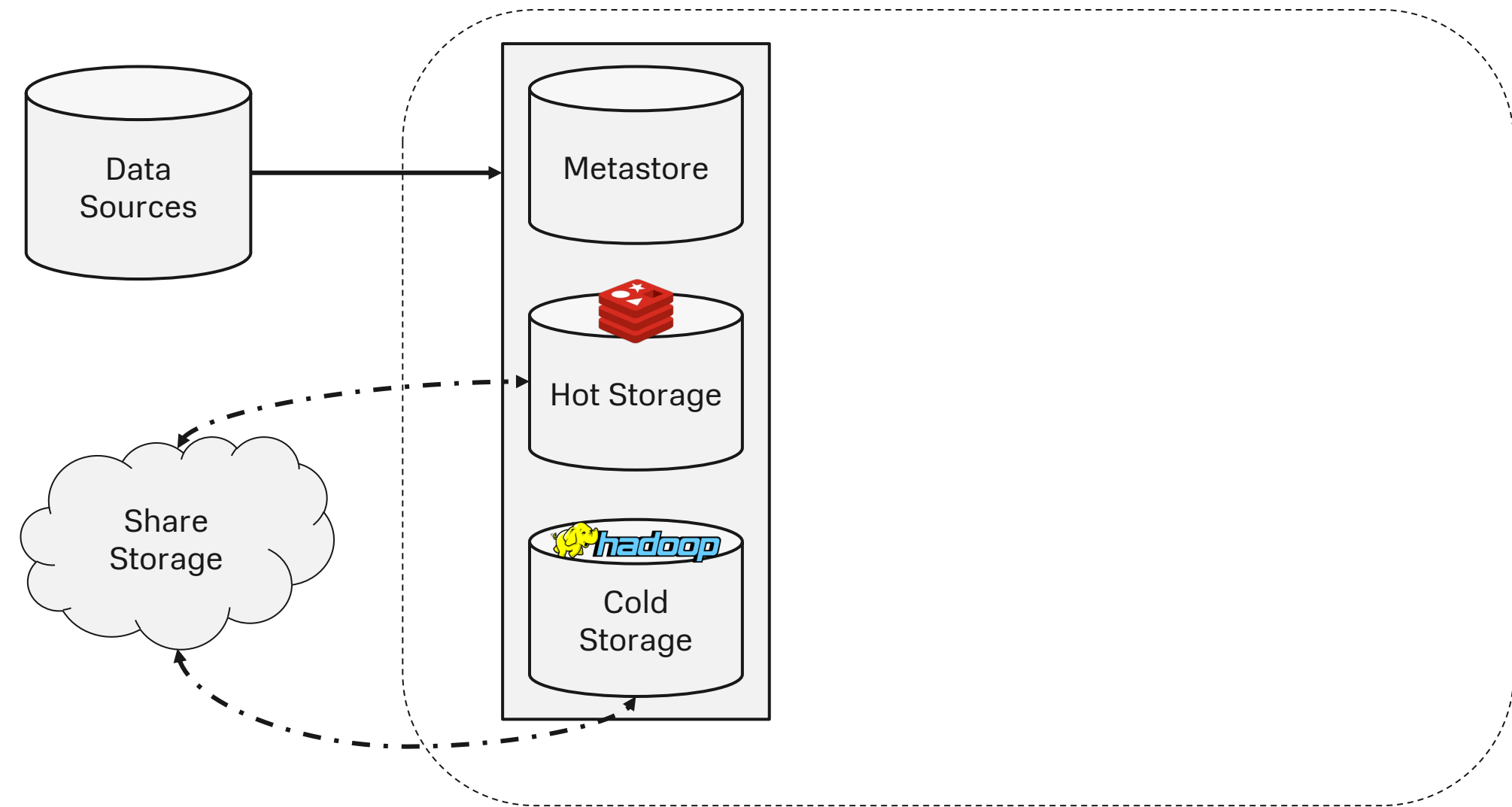
Проблемы:

-  Команды делают одно и то же
-  Долгий доступ
-  Отсутствие мониторинга

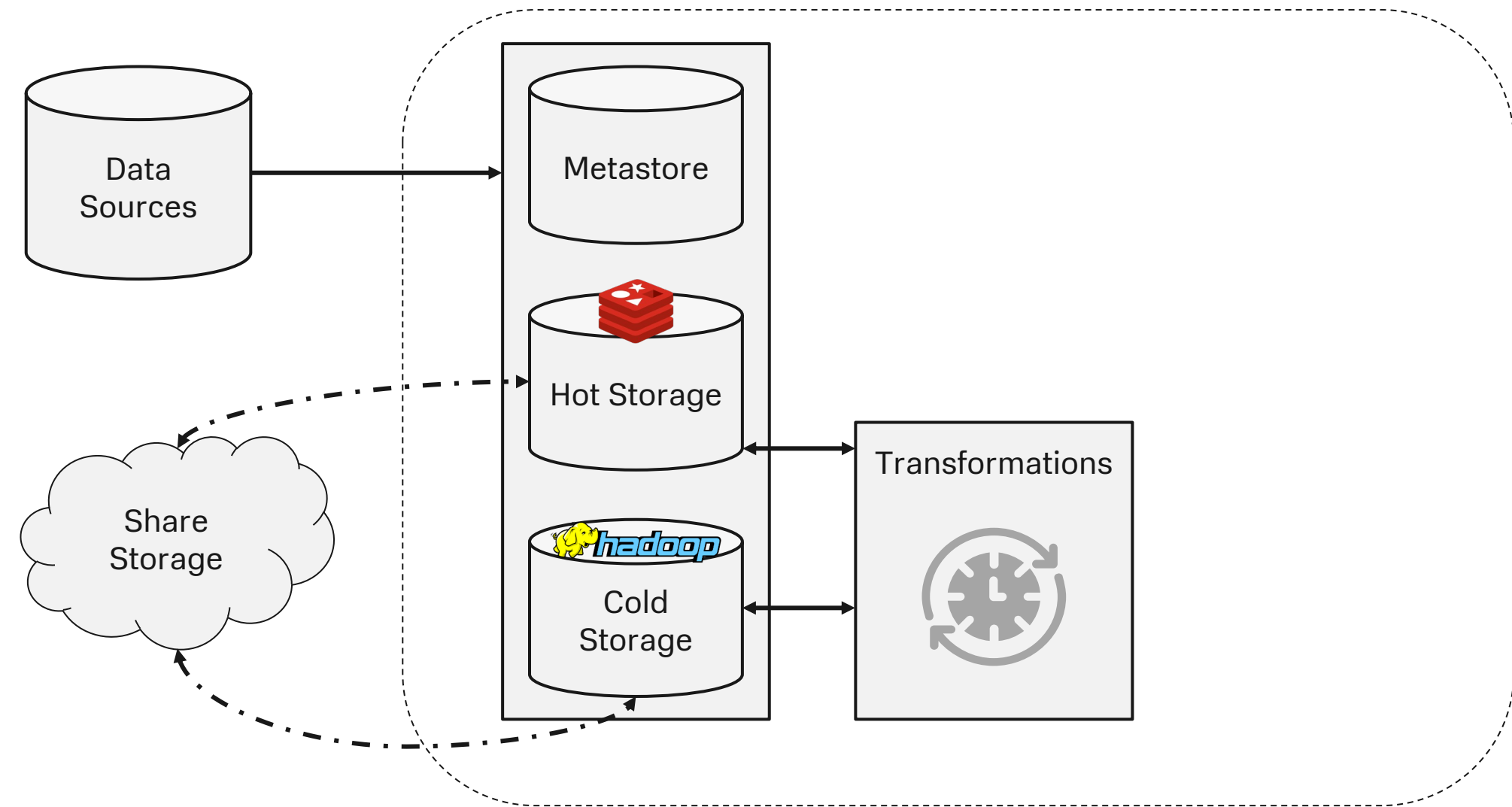
Feature Store



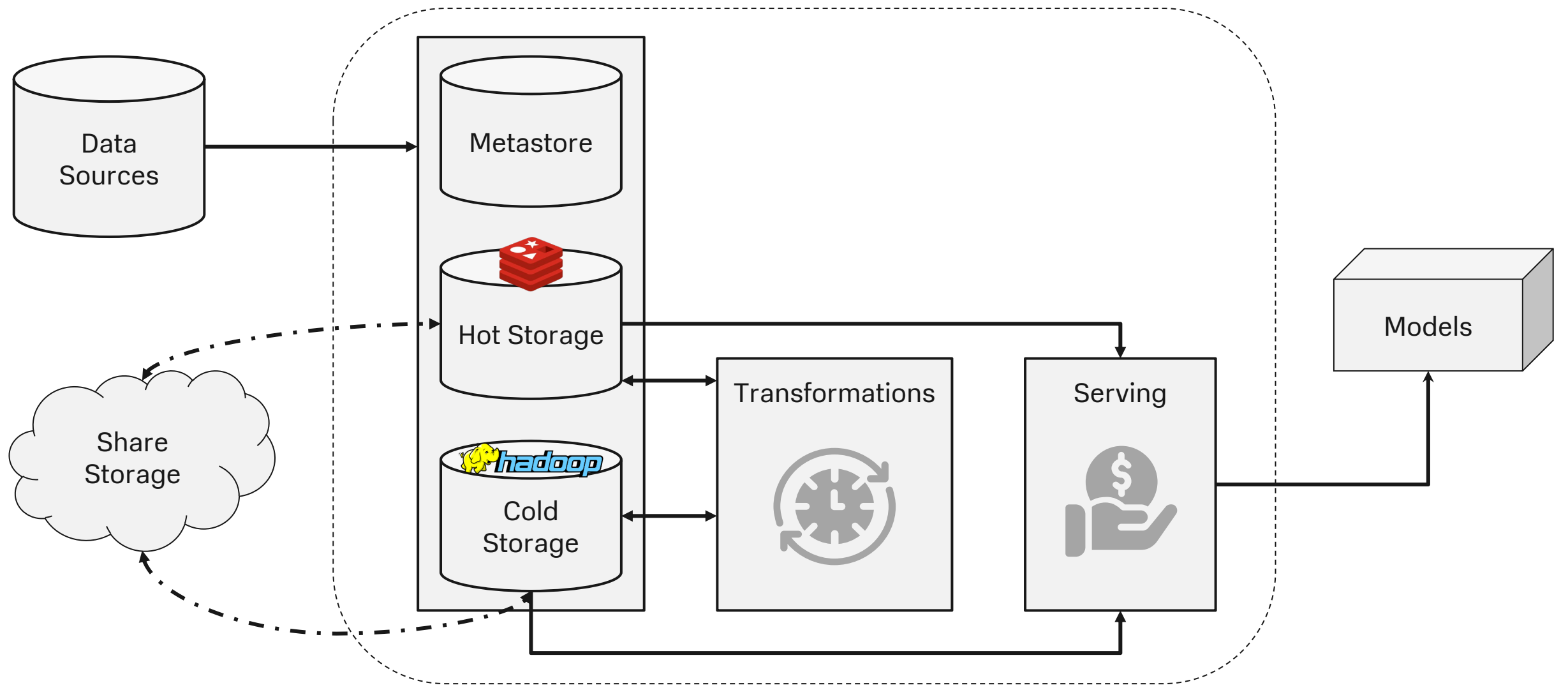
Feature Store



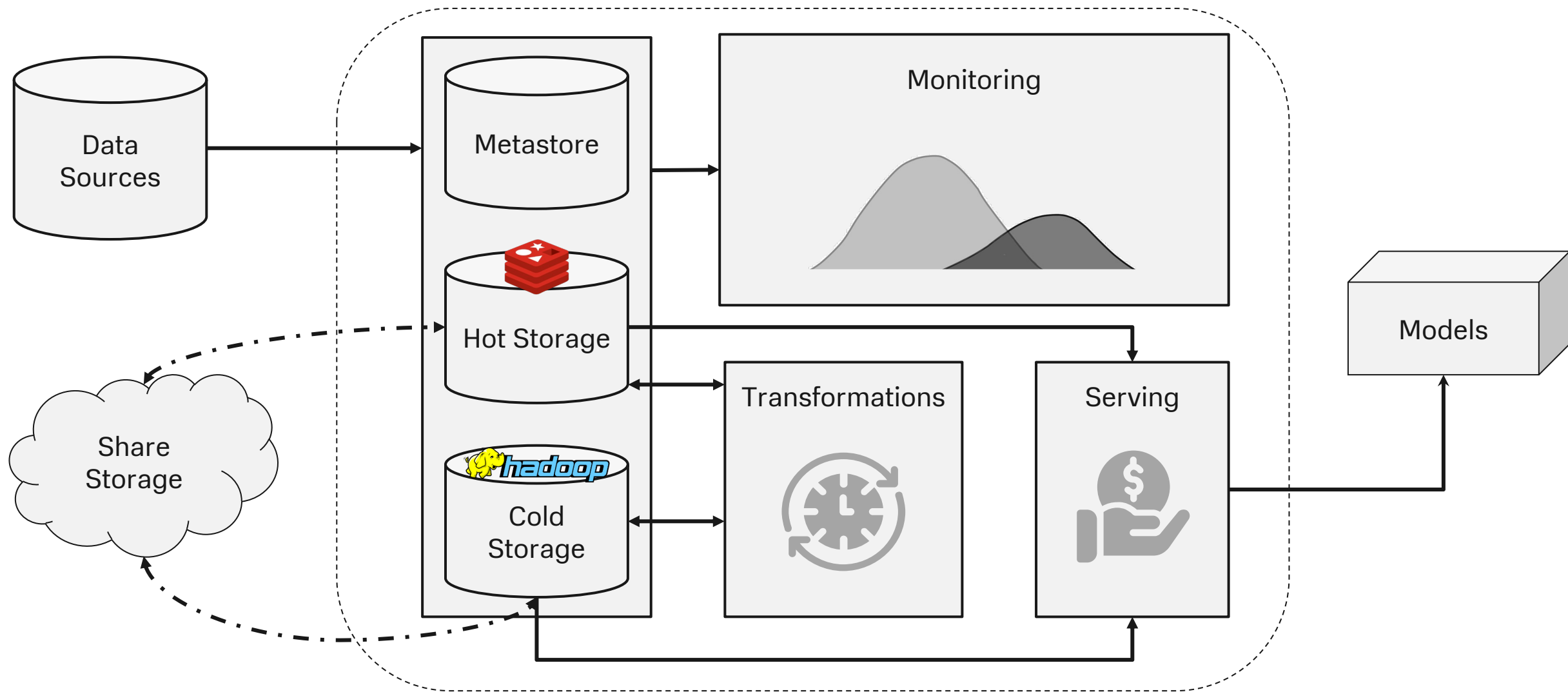
Feature Store



Feature Store



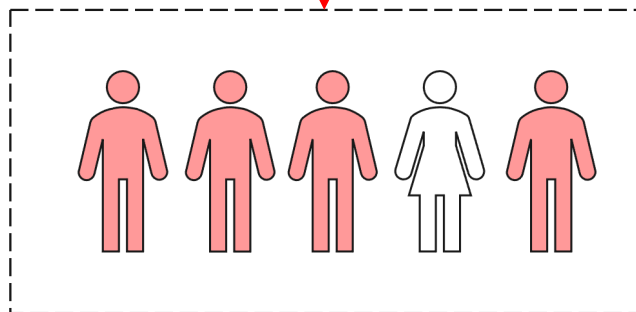
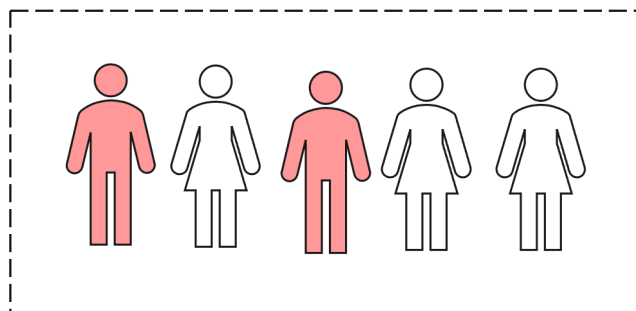
Feature Store



Динамика популяции



Изменения в составе



Динамика популяции



Изменения в составе



Естественные тренды



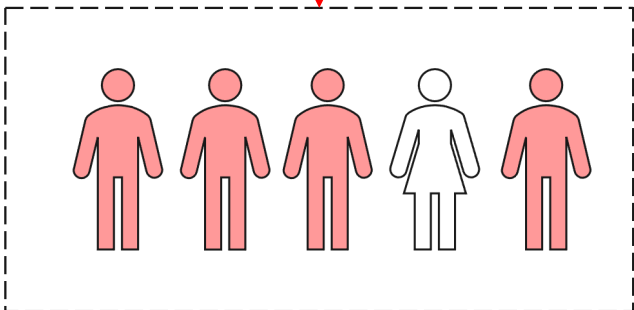
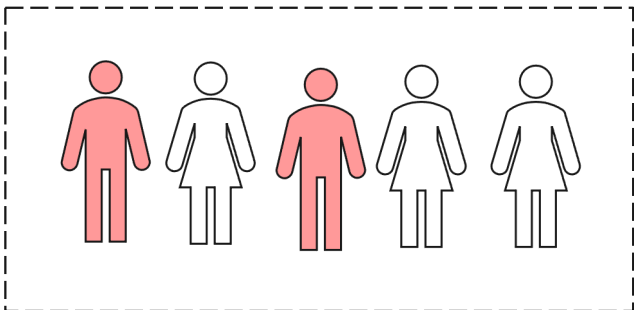
COVID-19



Праздники



Новые технологии



Динамика популяции



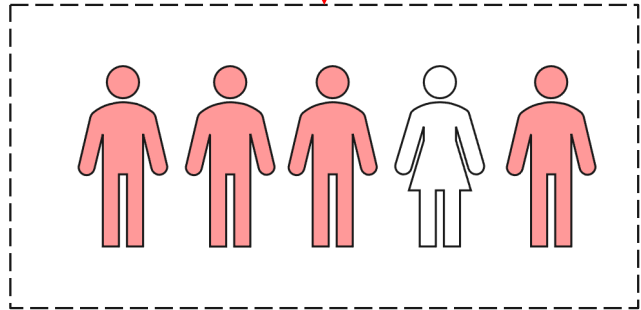
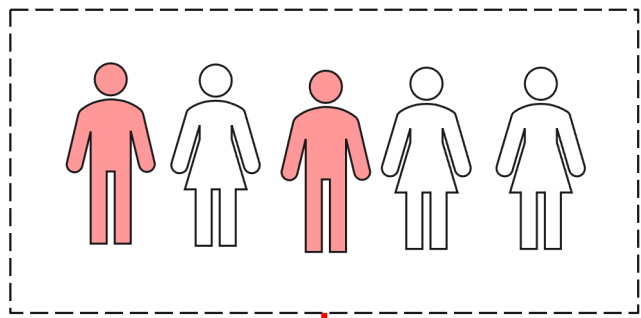
Изменения в составе



Естественные тренды



Ошибки при построении



COVID-19



Праздники



Новые технологии



Недоступность источника



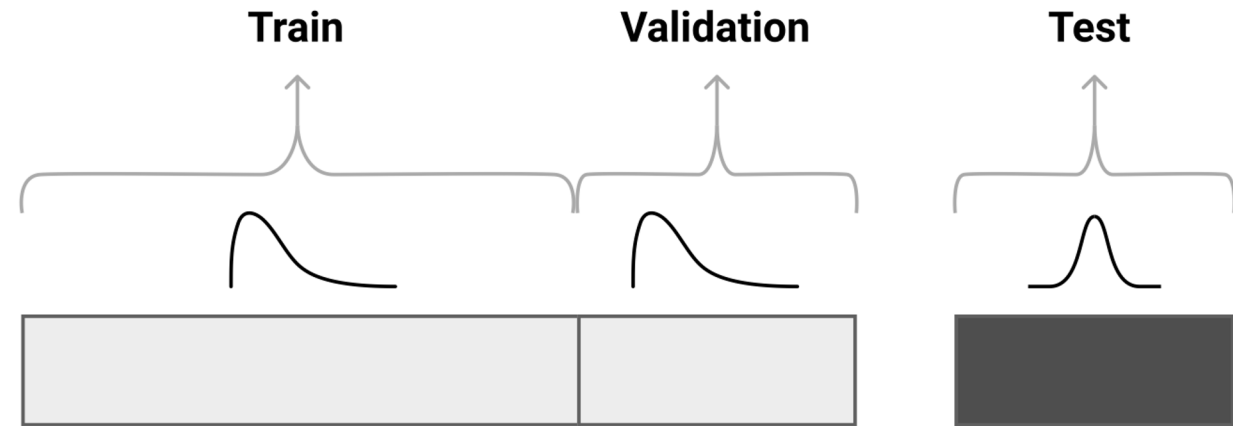
Данные собраны с ошибкой

Data Drift

Формальное определение

$$f : X \rightarrow Y$$

$$P(X, Y) \neq P_{\text{ref}}(X, Y)$$

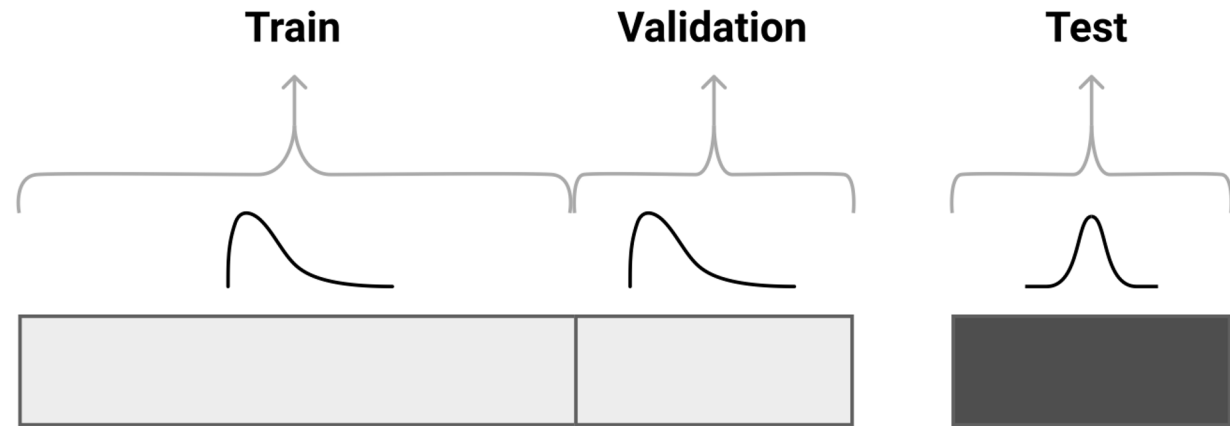


Data Drift

Формальное определение

$$f : X \rightarrow Y$$

$$P(X, Y) \neq P_{\text{ref}}(X, Y)$$



Мониторинг – сравниваем новый поток с тем, на котором обучалась модель.

Цель мониторинга – детекция оснований для изменения качества модели.

Постановка задачи

Задача – разработка автоматизированного решения для DQ мониторинга в Feature Store:



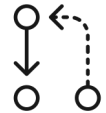
Детекция дрейфа

Постановка задачи

Задача – разработка автоматизированного решения для DQ мониторинга в Feature Store:



Детекция дрейфа



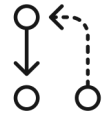
Построение DQ пайплайнов

Постановка задачи

Задача – разработка автоматизированного решения для DQ мониторинга в Feature Store:



Детекция дрейфа



Построение DQ пайплайнов



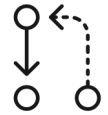
Обработка инцидентов

Постановка задачи

Задача – разработка автоматизированного решения для DQ мониторинга в Feature Store:



Детекция дрейфа



Построение DQ пайплайнов



Обработка инцидентов



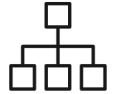
Анализ и исправление исторических данных

| 2

Существующие решения

Существующие решения

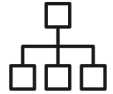
Основные требования:



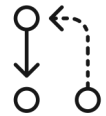
Методы работают в распределенном стеке

Существующие решения

Основные требования:



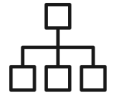
Методы работают в распределенном стеке



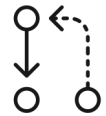
Подходит для ML мониторинга

Существующие решения

Основные требования:



Методы работают в распределенном стеке



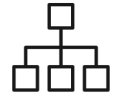
Подходит для ML мониторинга



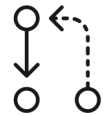
Подходит для DQ мониторинга признакового пространства

Существующие решения

Основные требования:



Методы работают в распределенном стеке



Подходит для ML мониторинга



Подходит для DQ мониторинга признакового пространства



Интегрируется с нашим DQ-метастором

Существующие решения

PyDeequ – Unit Tests for Data



Существующие решения

PyDeequ – Unit Tests for Data

- ✓ Работает в распределенном стеке
- ✓ Готовое решения для мониторинга признакового пространства



Существующие решения

PyDeequ – Unit Tests for Data

- ✓ Работает в распределенном стеке
- ✓ Готовое решения для мониторинга признакового пространства
- ✗ Нужно добавлять функционал для мониторинга целевых признаков
- ✗ Потребуется дополнительные ресурсы на интеграцию с нашим DQ-метастором



Существующие решения

Evidently и Deepchecks



Существующие решения

Evidently и Deepchecks



Широкий функционал для ML мониторинга



Существующие решения

Evidently и Deepchecks



Широкий функционал для ML мониторинга



Потребуется ресурс на адаптацию к распределенному стеку












Потребуется ресурс на интеграцию с нашим DQ-метастором



Существующие решения

Сравнение

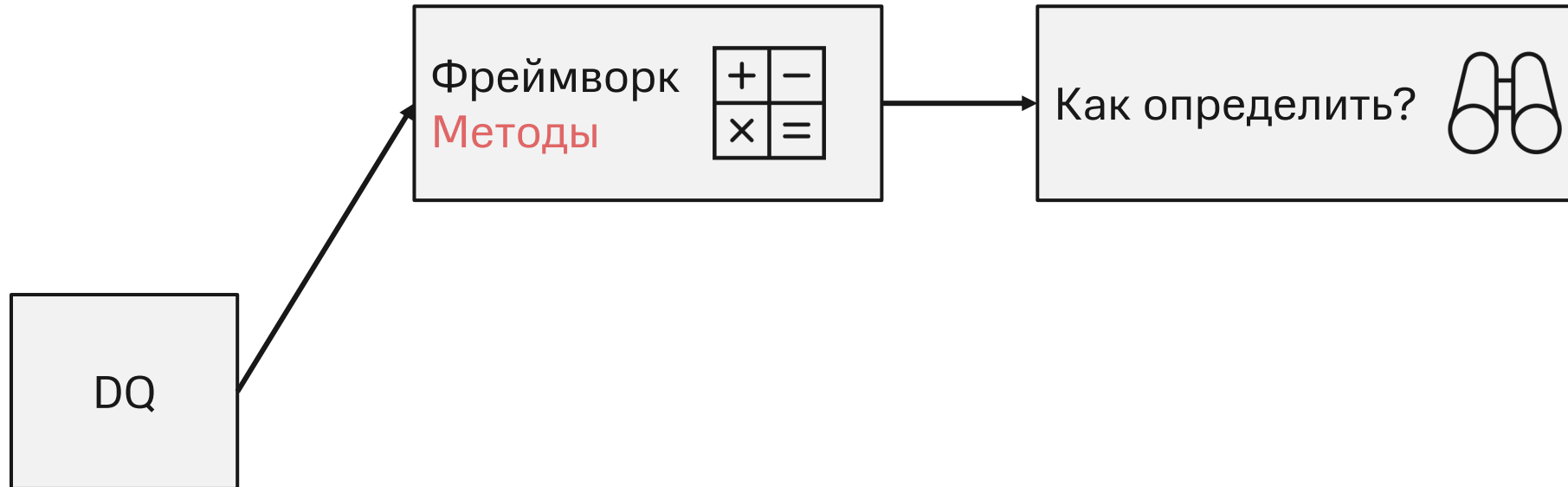
	Spark	ML Monitoring	DQ Monitoring
PyDeequ			
Deepchecks			
Evidently			

| 3

Реализация

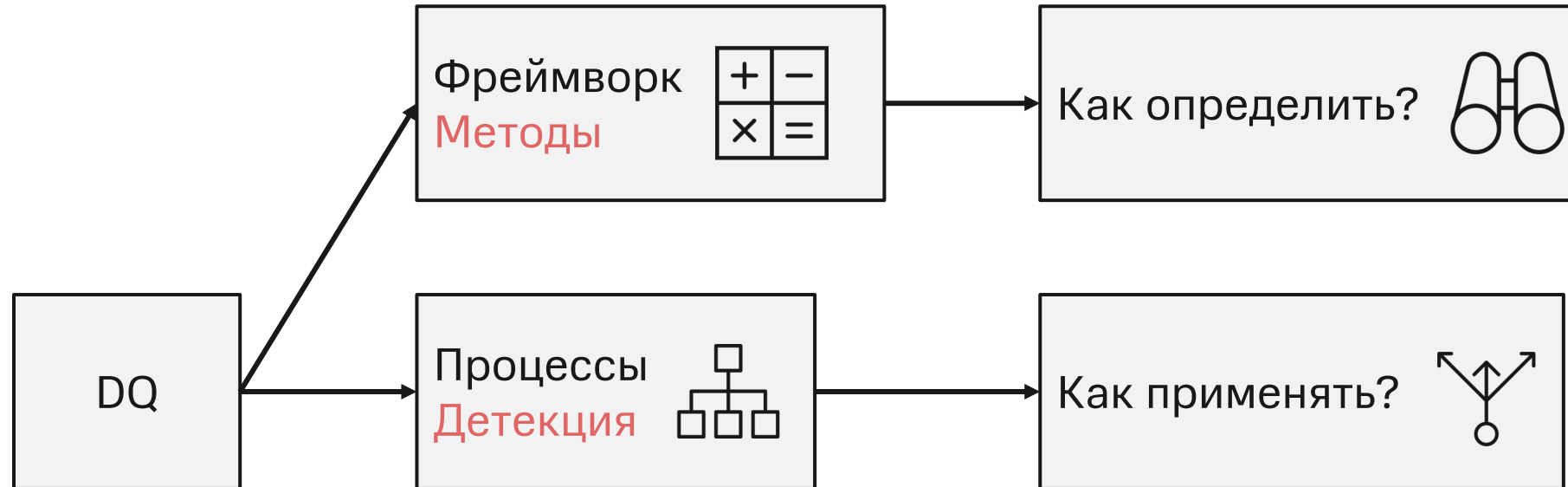
Реализация

Структура решения

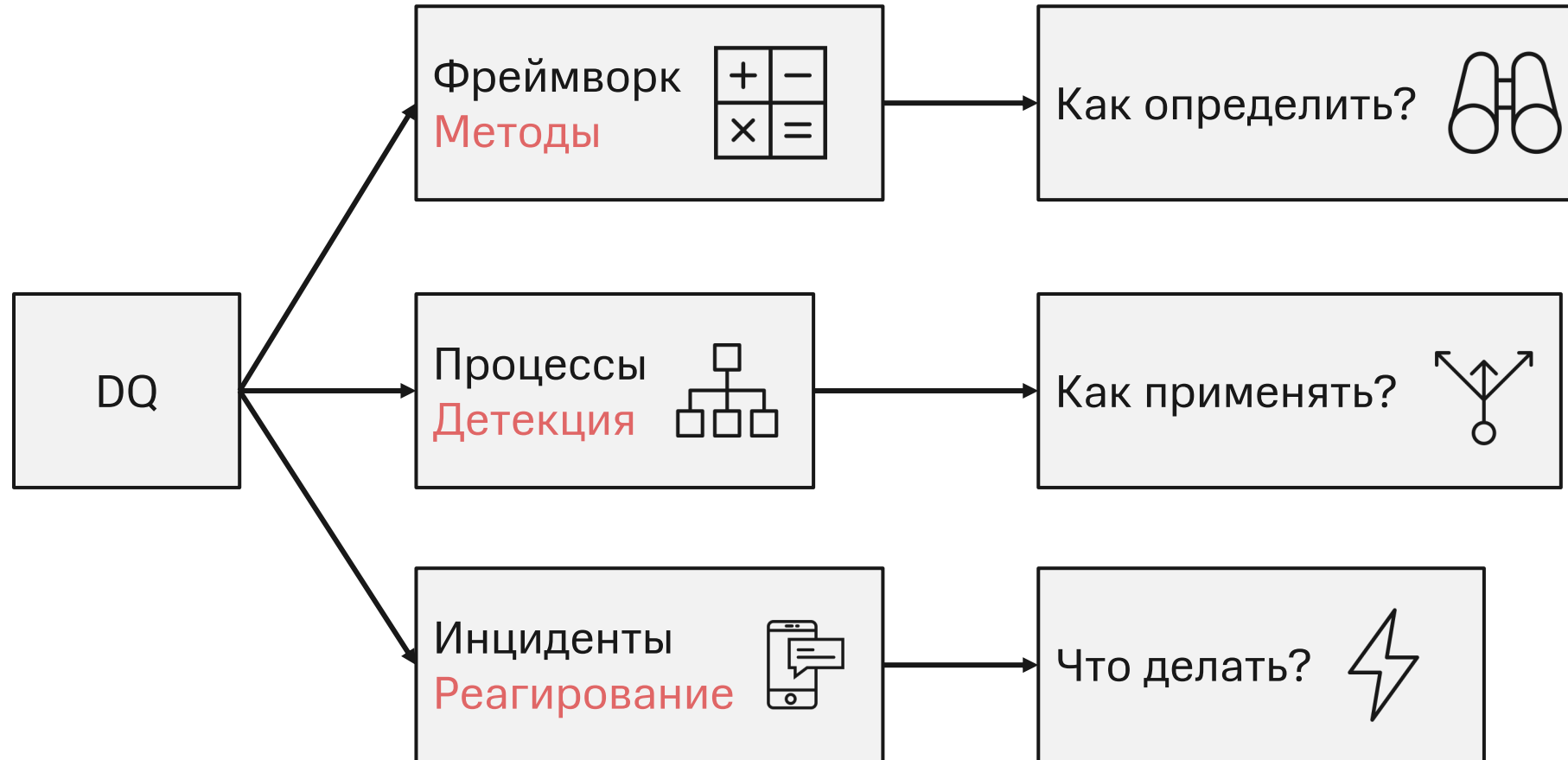


Реализация

Структура решения



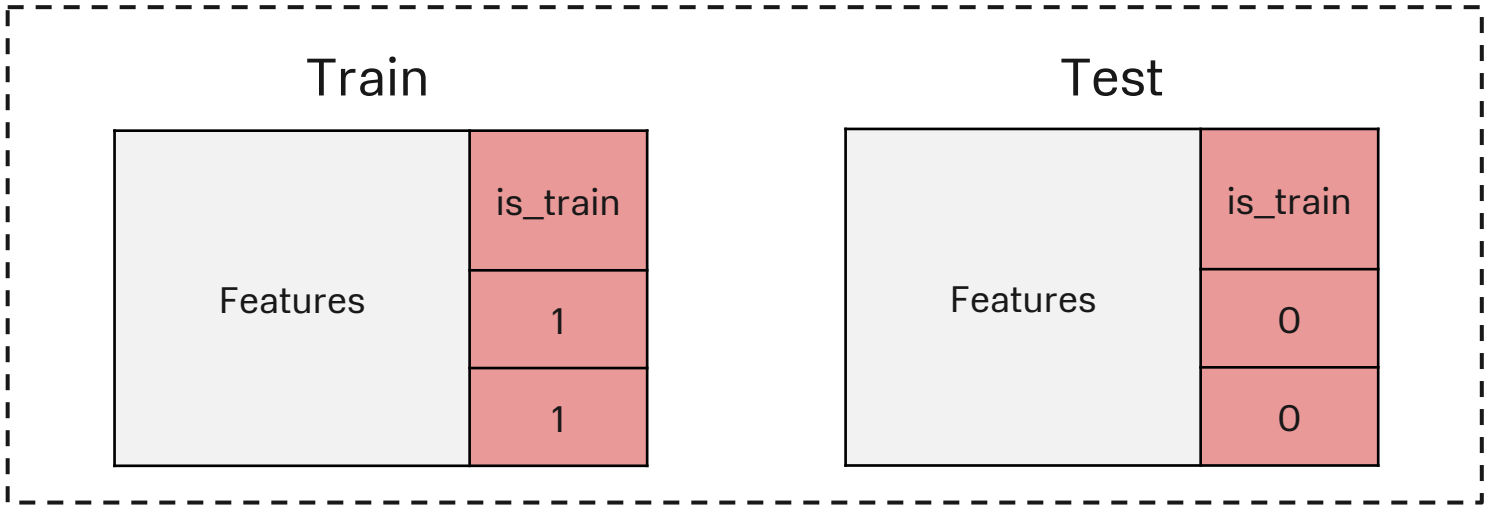
Структура решения



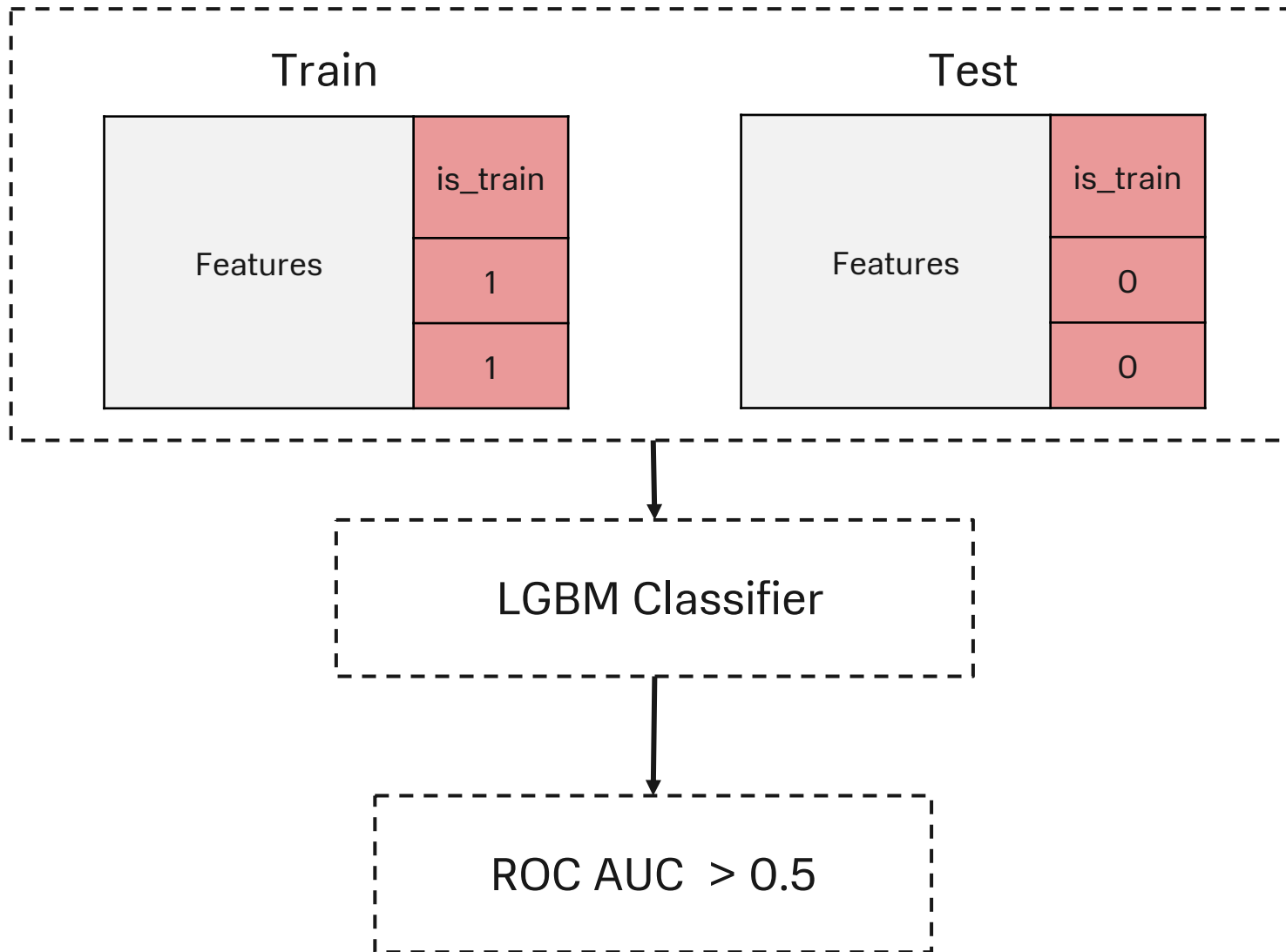
| 3.1

Реализация Методы

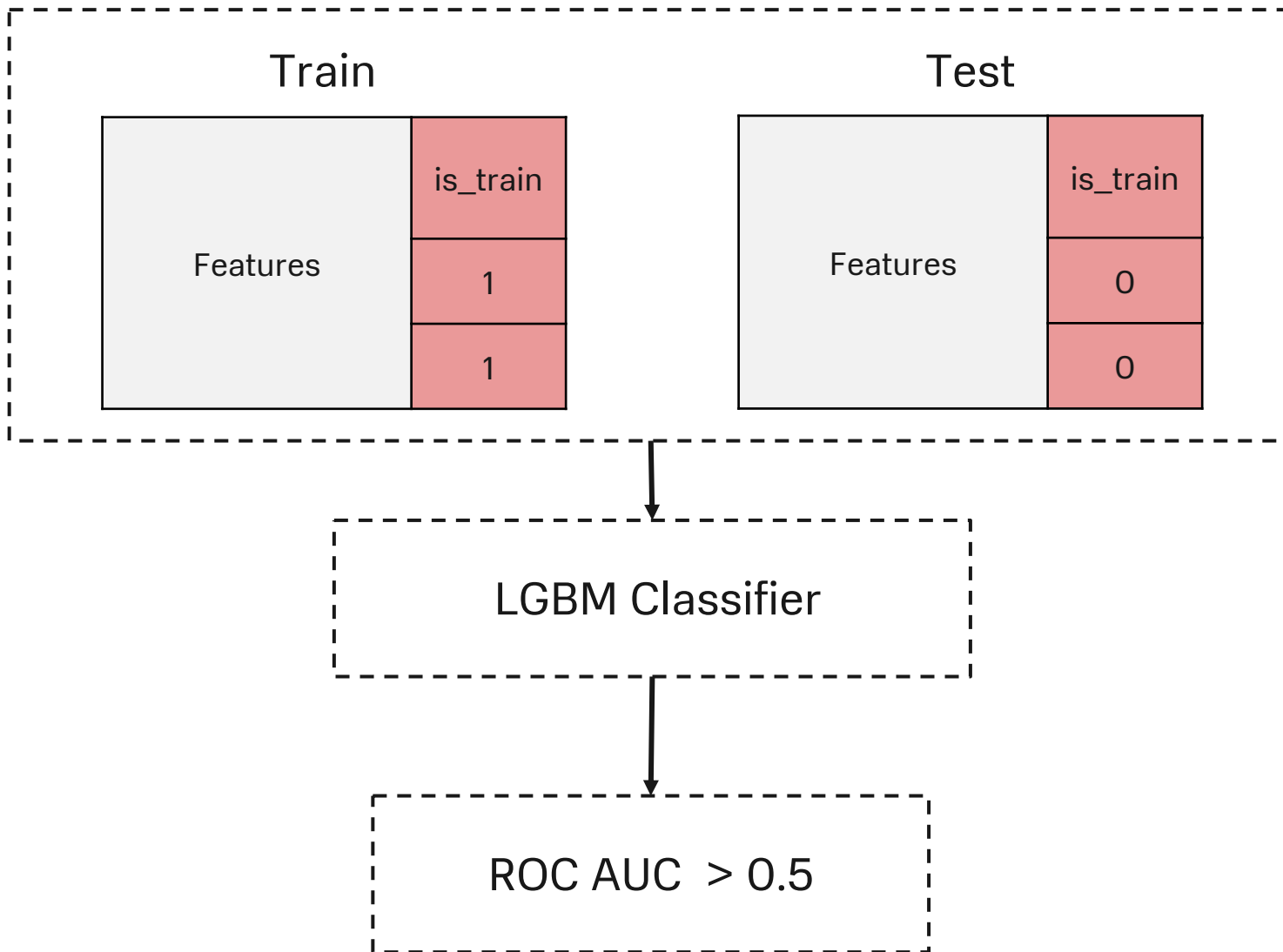
Adversarial Validation



Adversarial Validation



Adversarial Validation

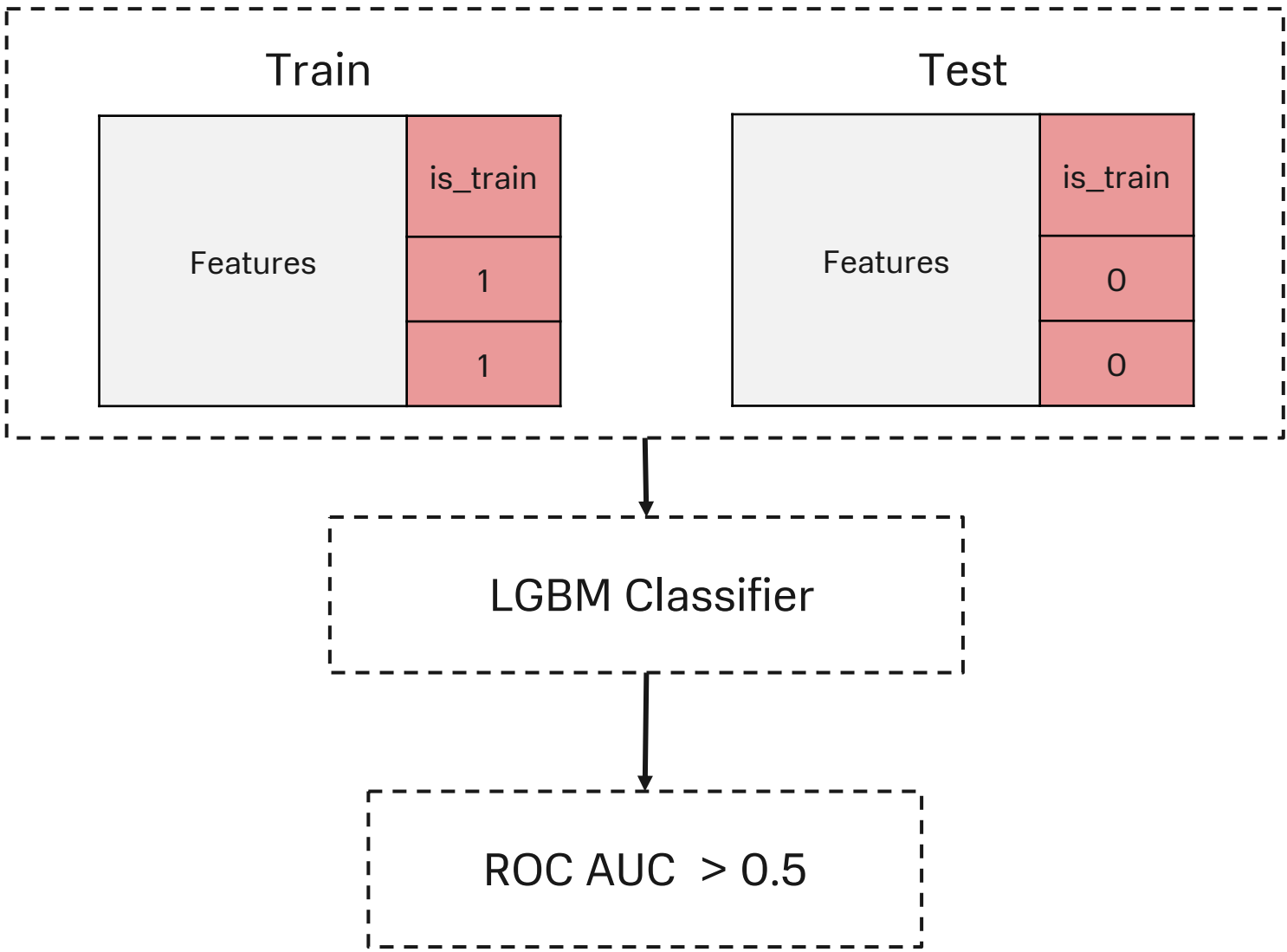


Рассматриваем совокупное признаковое пространство



Можно использовать для отбора признаков на обучении

Adversarial Validation



Рассматриваем совокупное признаковое пространство



Можно использовать для отбора признаков на обучении

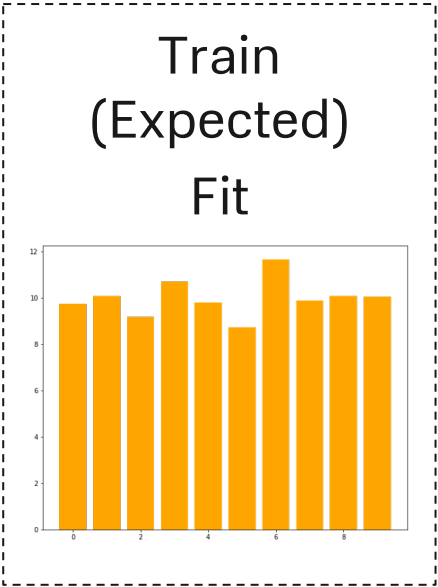


Обучать сильную модель долго

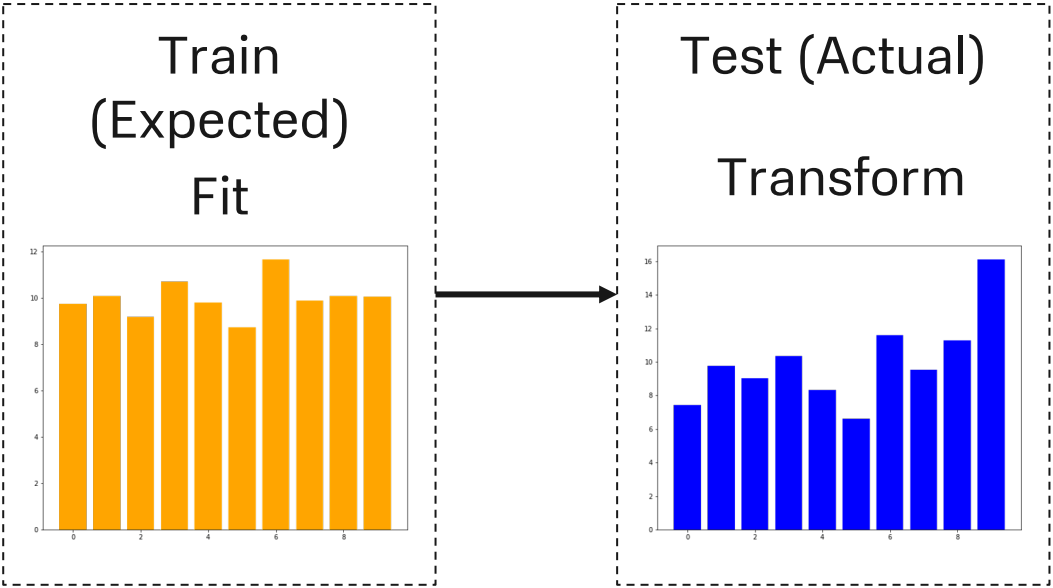


Слабо интерпретируемая метрика

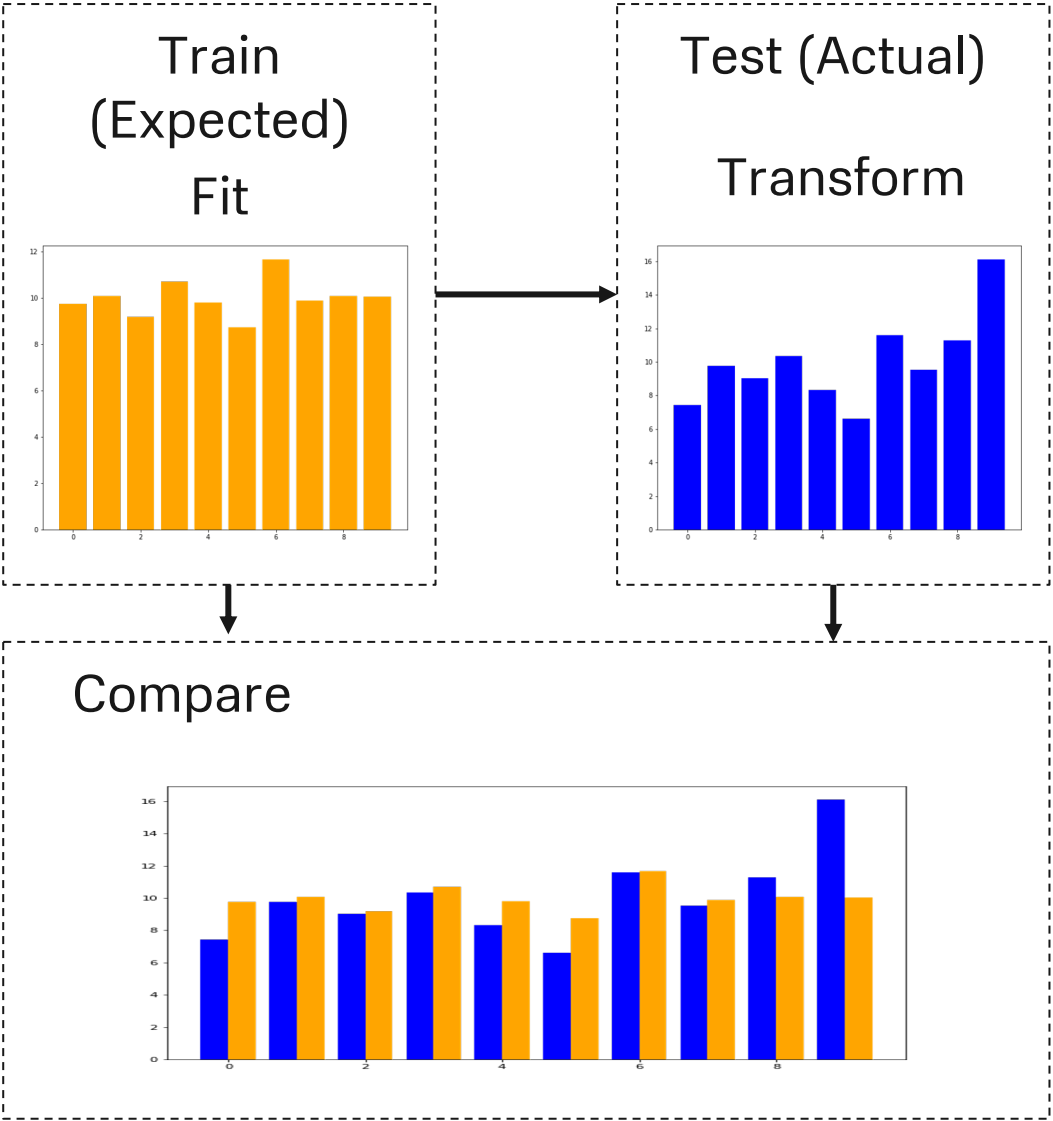
Population Stability Index



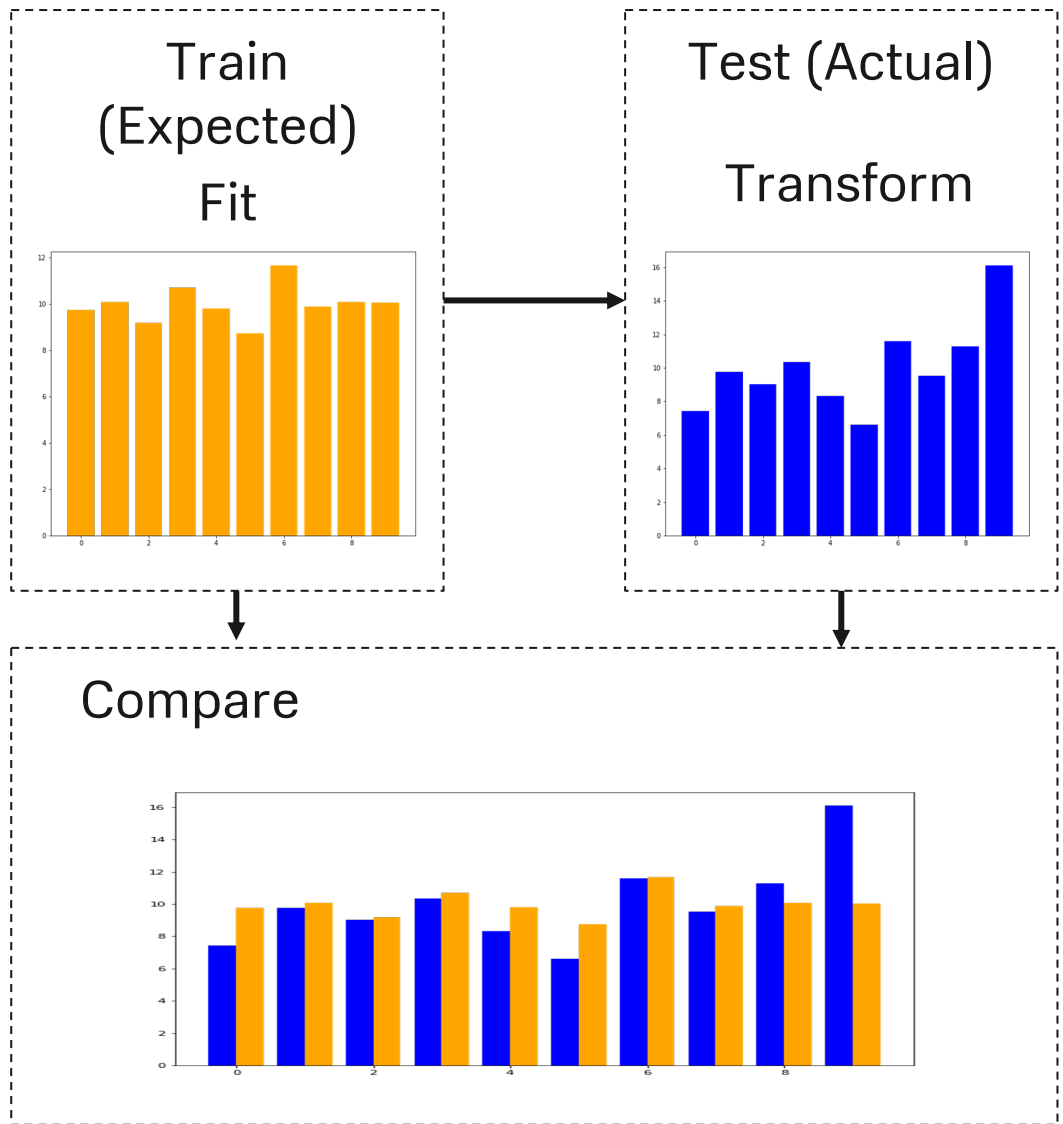
Population Stability Index



Population Stability Index



Population Stability Index



Легко считается распределенно

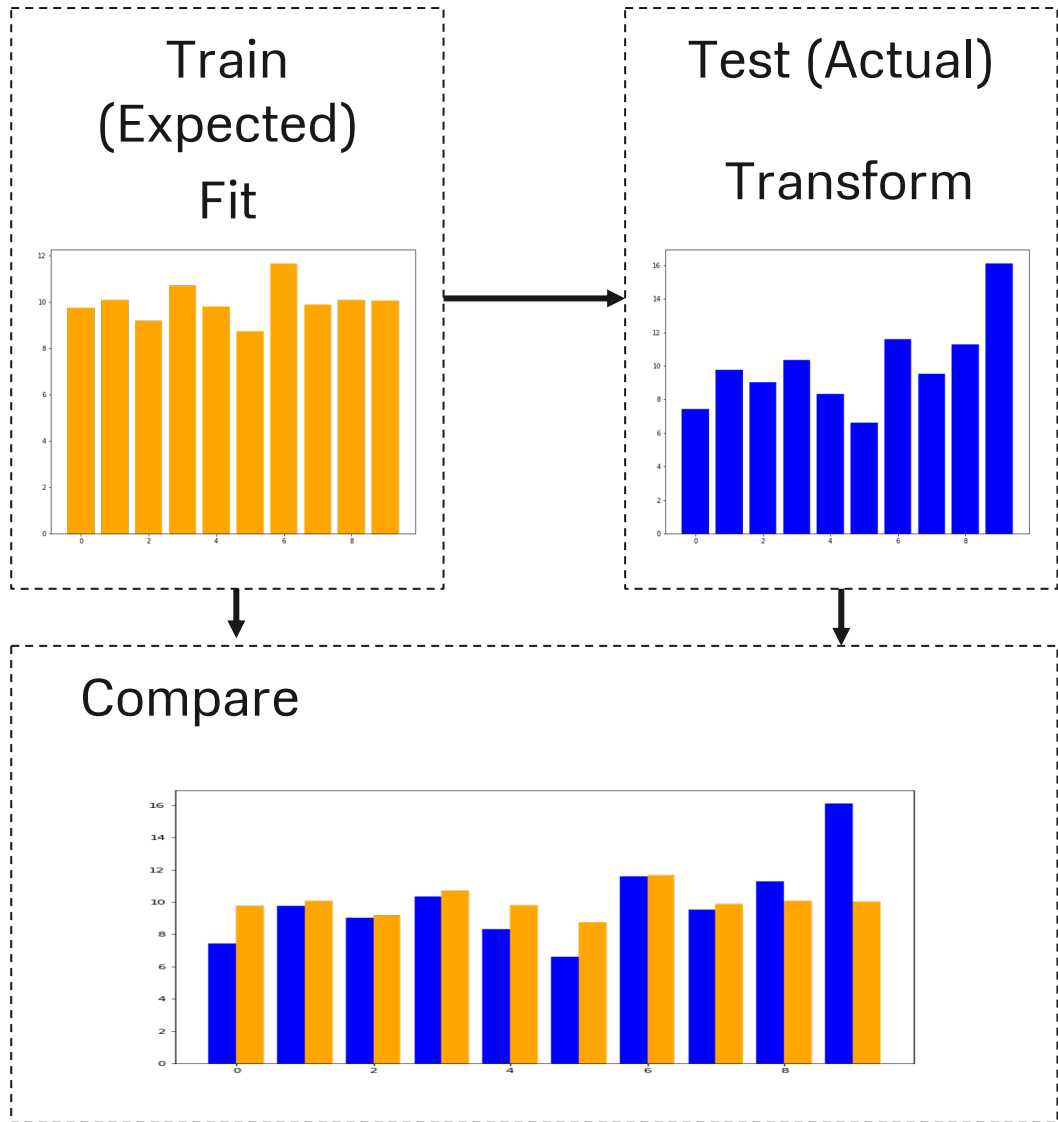


Потенциал оптимизации



Интерпретируемые значения

Population Stability Index



Легко считается распределенно



Потенциал оптимизации



Интерпретируемые значения



Рассматриваем маргинальные распределения

| 3.4

Реализация Процессы

Процессы

Классификация

Задача – построить DQ пайплайны вокруг следующих сущностей:



Группа признаков в Feature Store

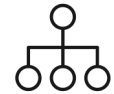
Процессы

Классификация

Задача – построить DQ пайплайны вокруг следующих сущностей:



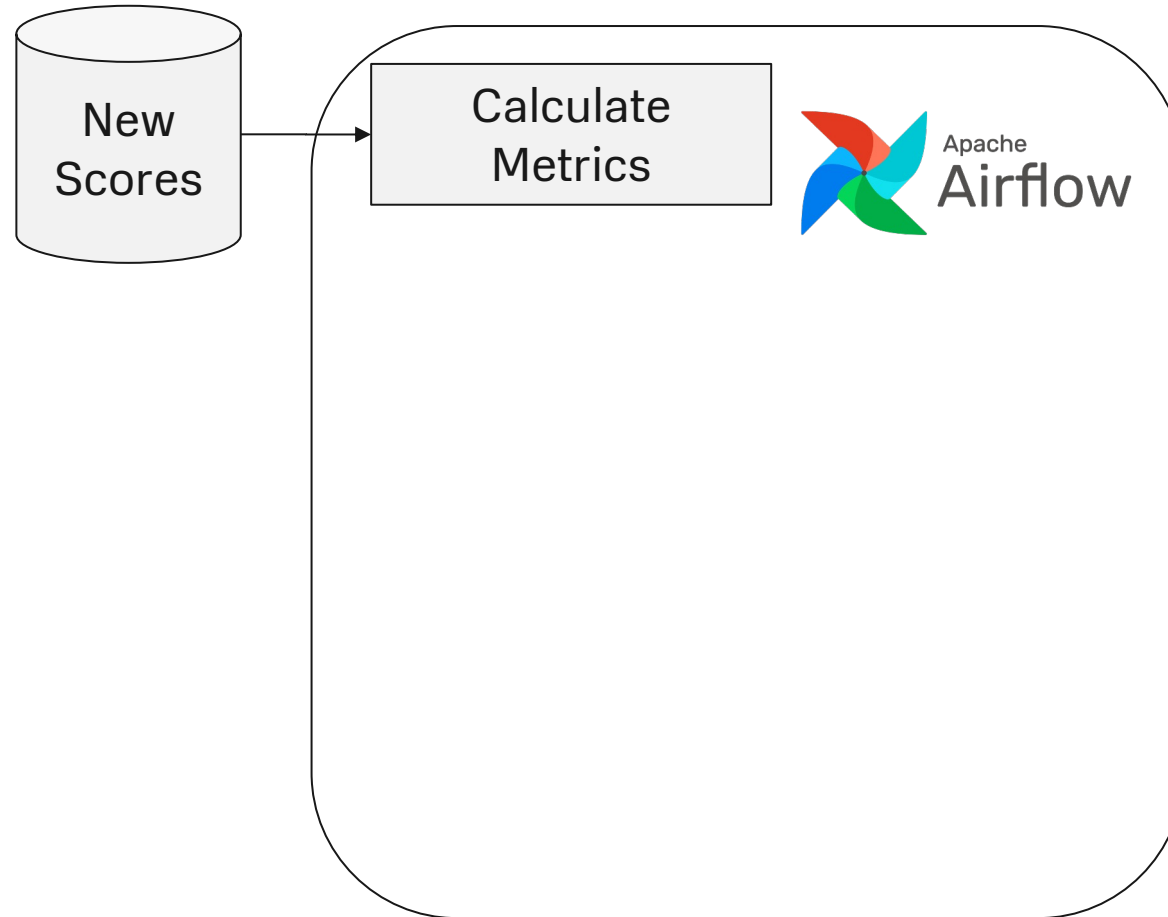
Группа признаков в Feature Store



ML модель в продакшене

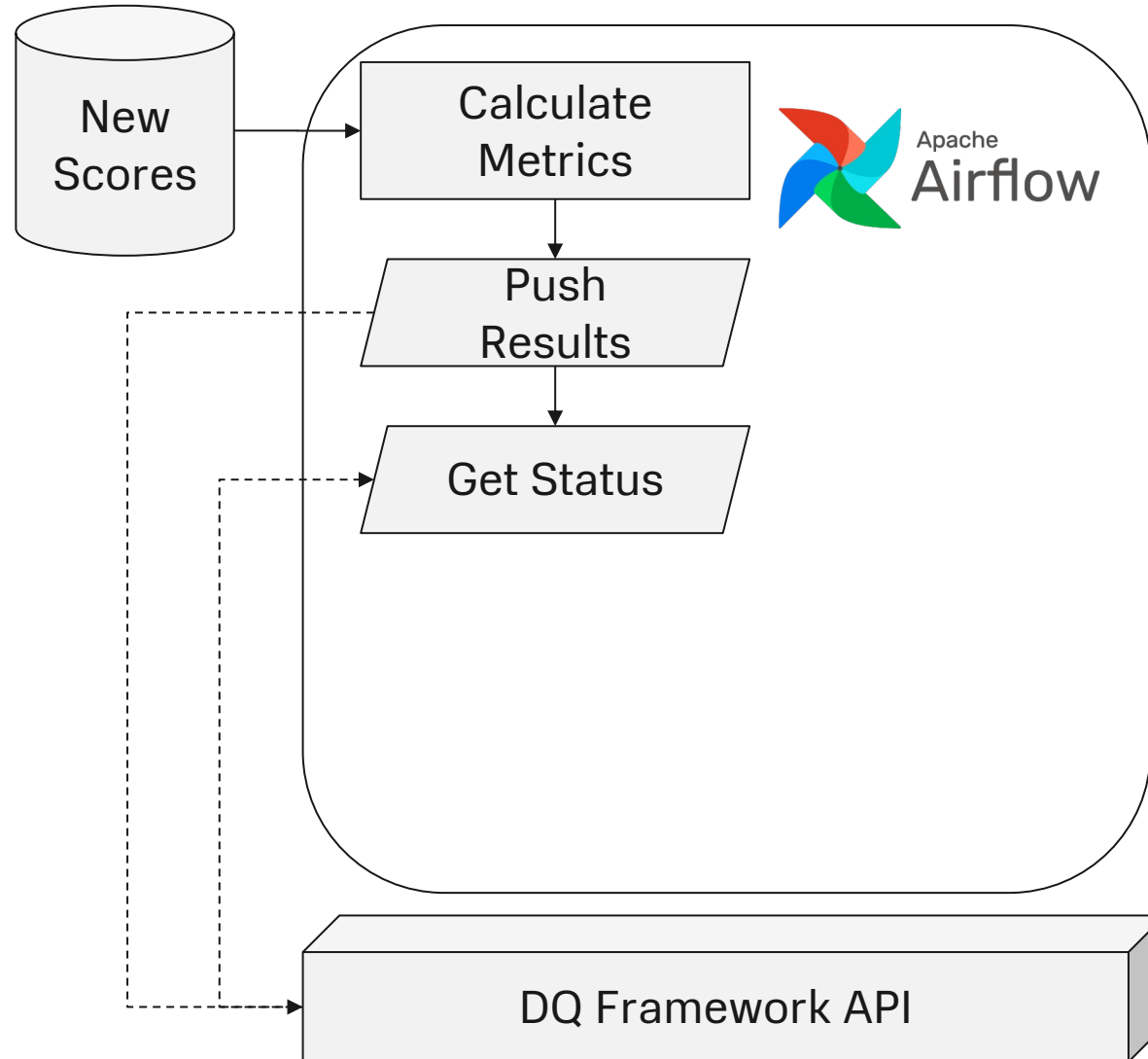
Процессы

ML модель в продакшене



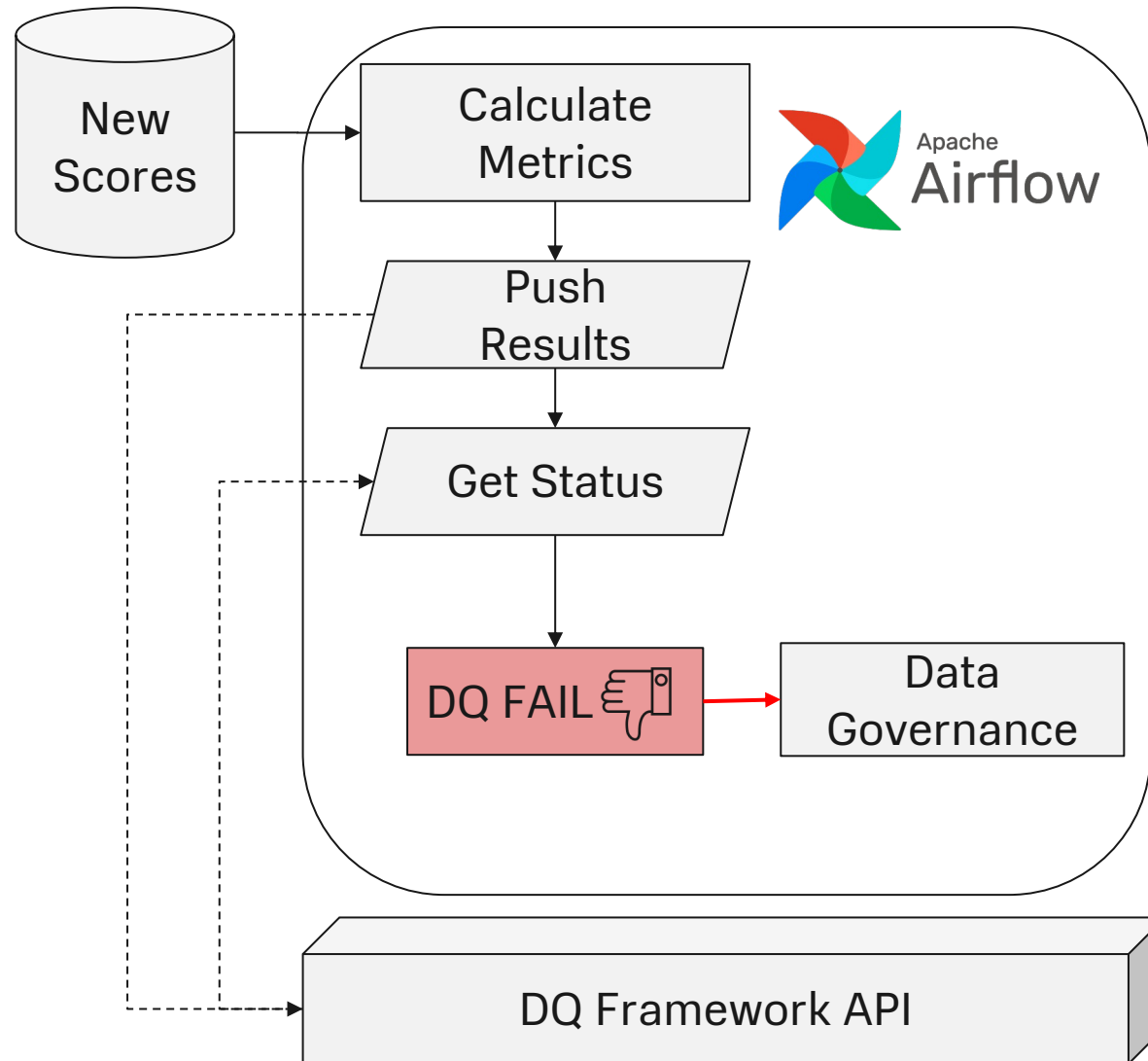
Процессы

ML модель в продакшене



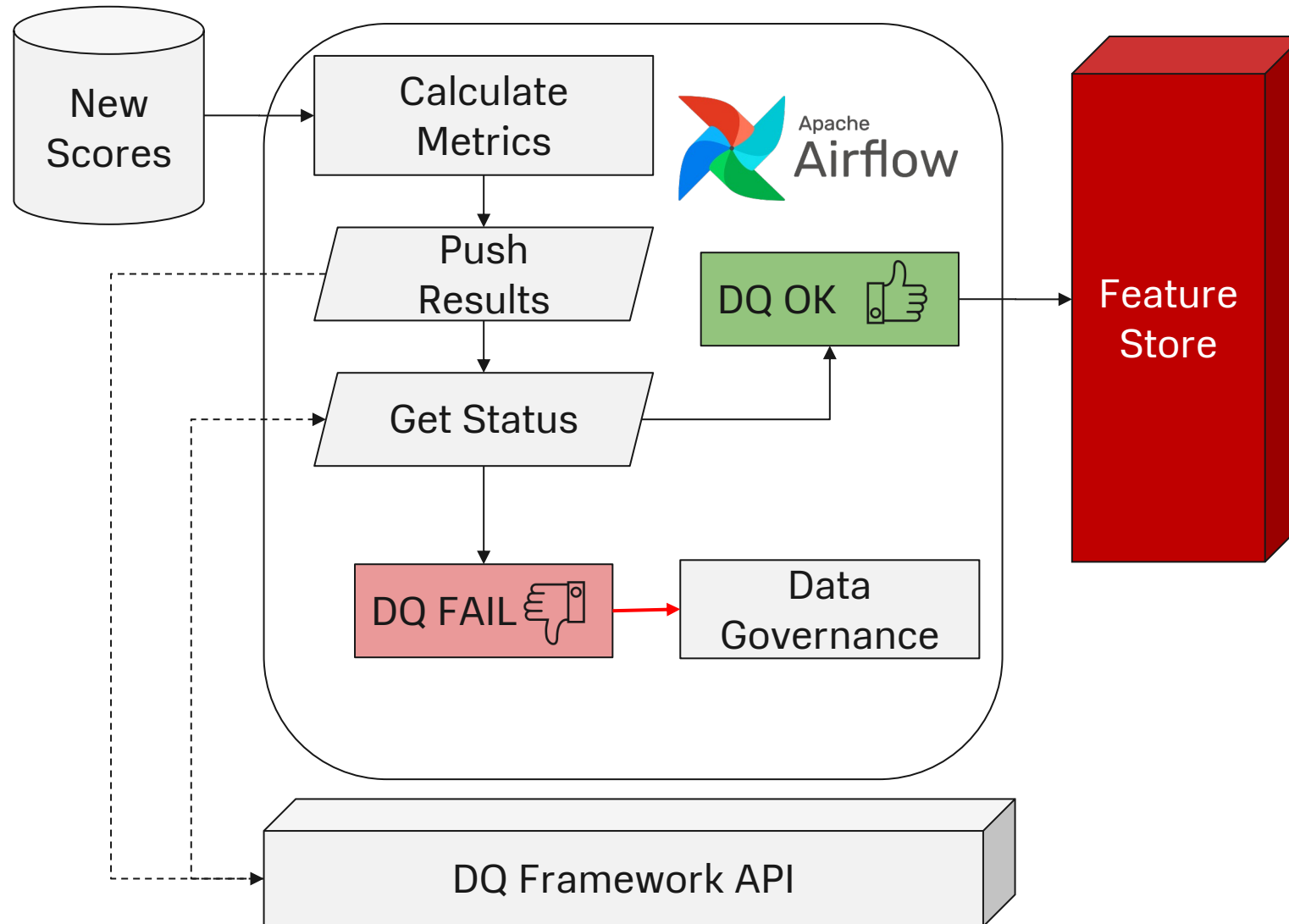
Процессы

ML модель в продакшене



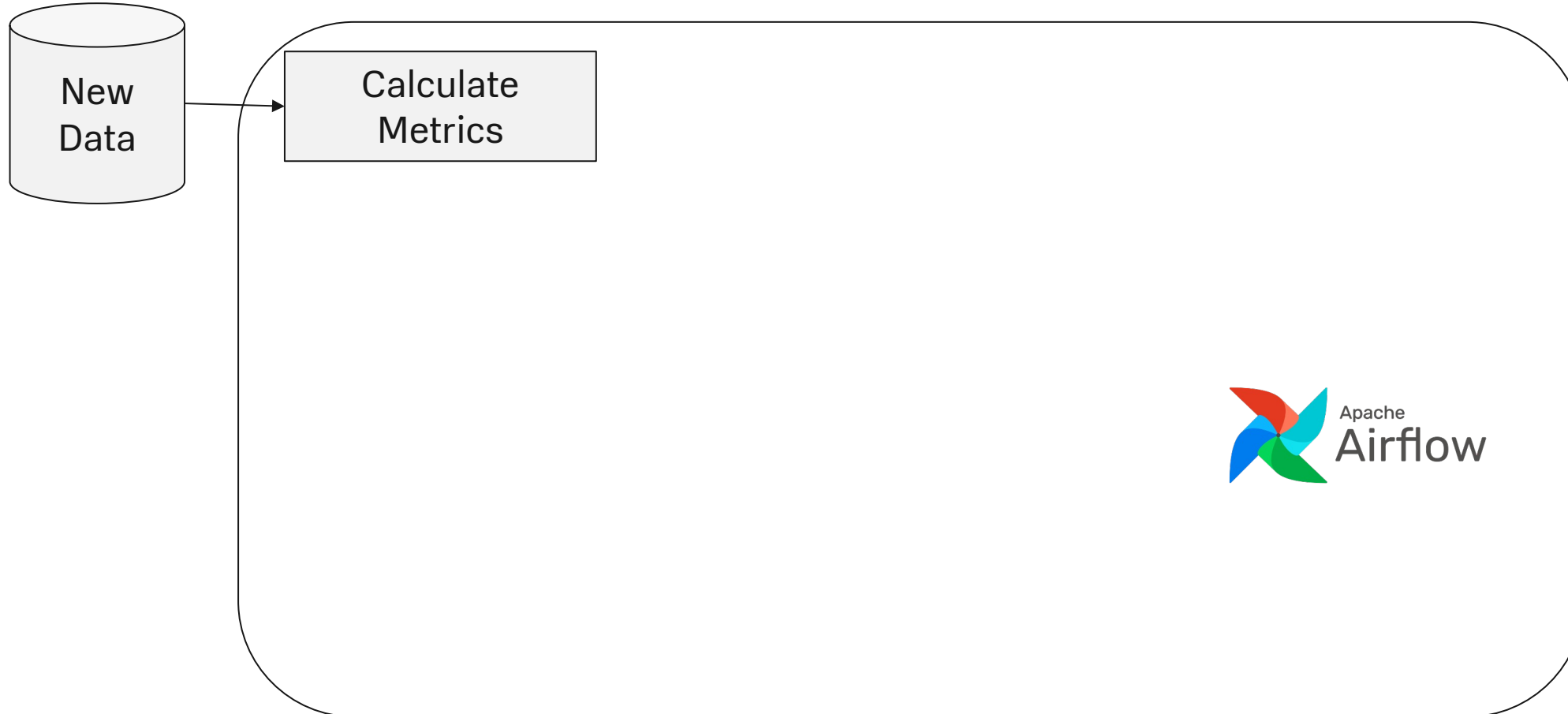
Процессы

ML модель в продакшене

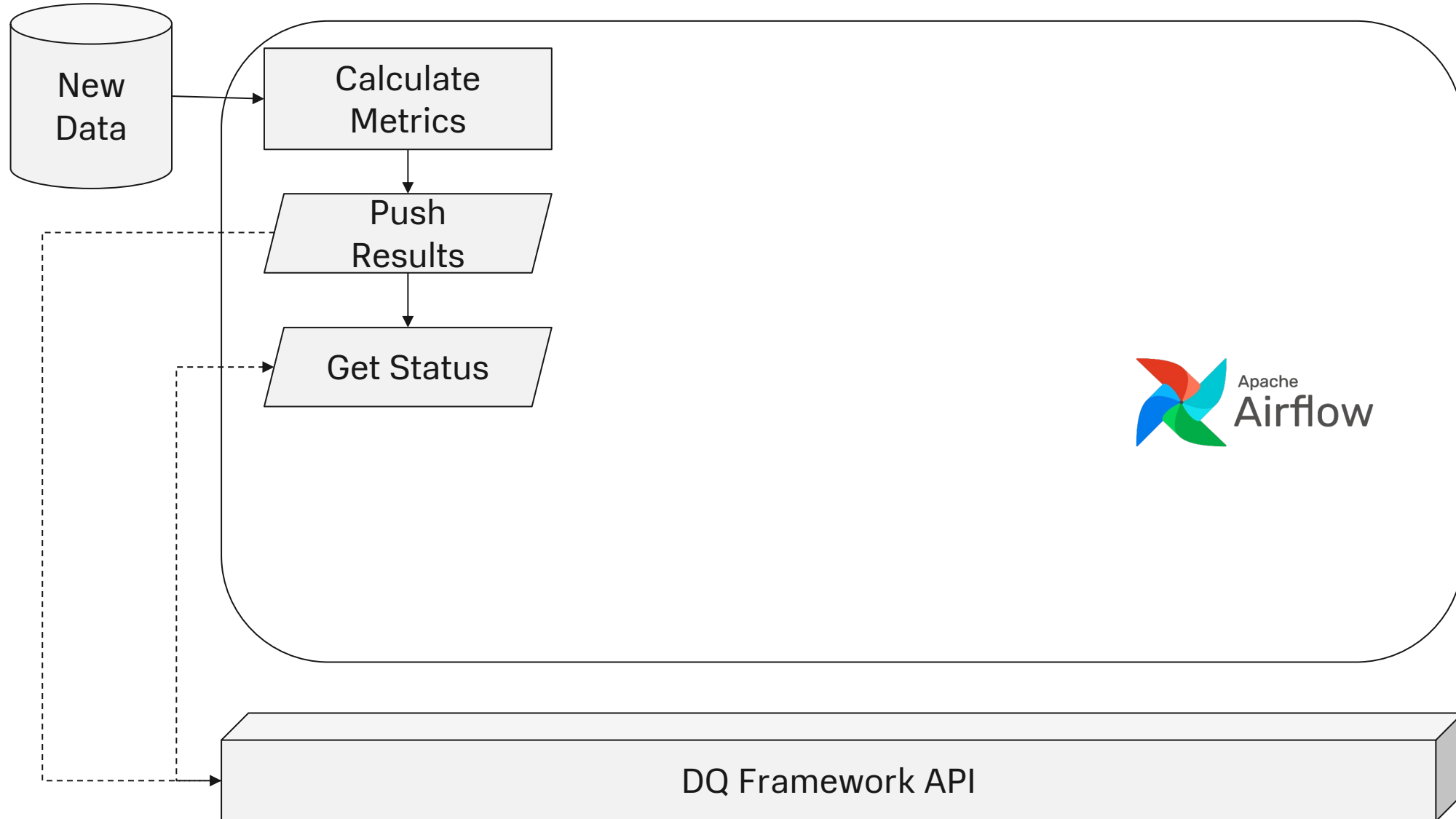


Процессы

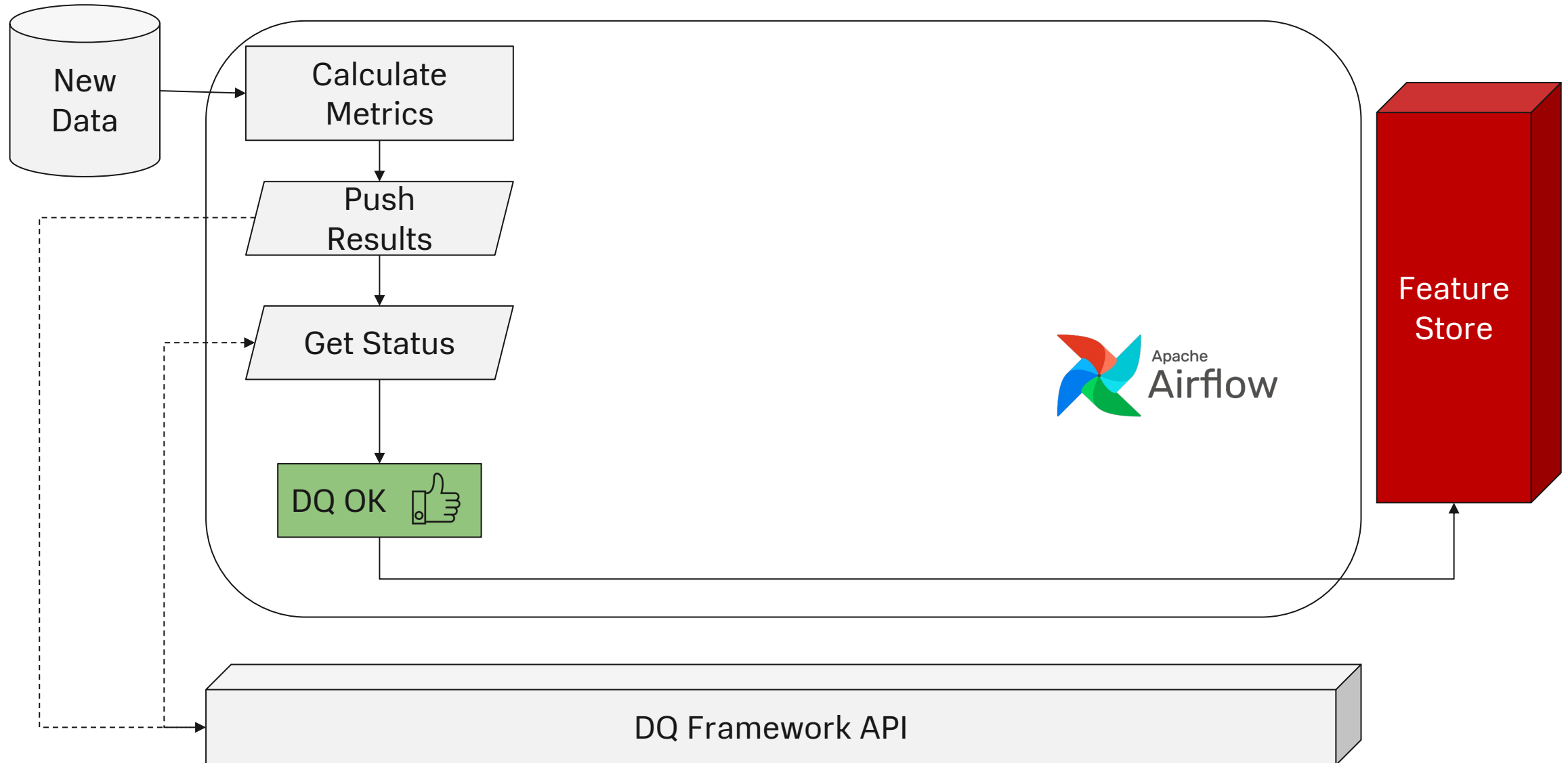
Группа признаков в Feature Store



Группа признаков в Feature Store

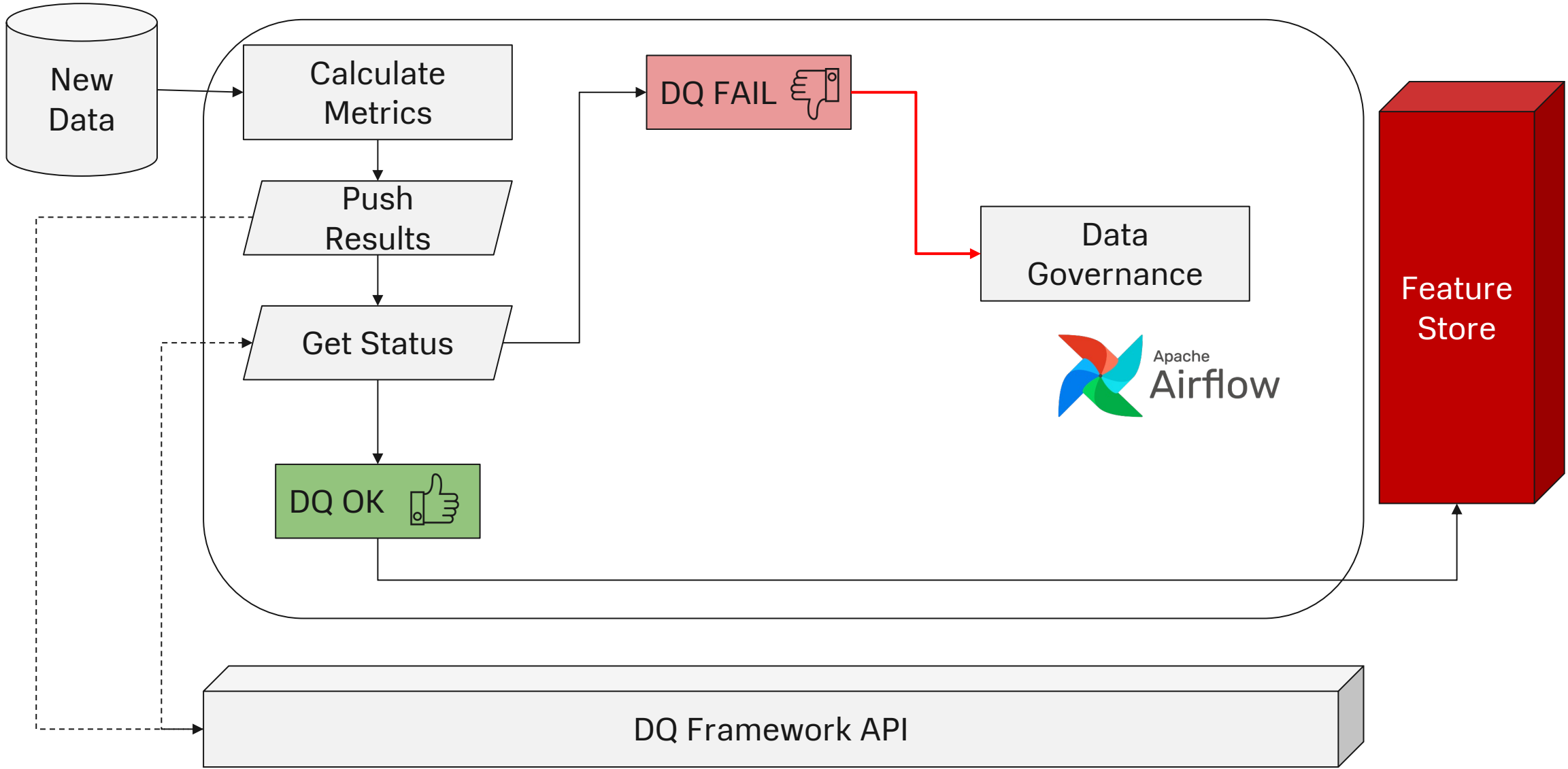


Группа признаков в Feature Store



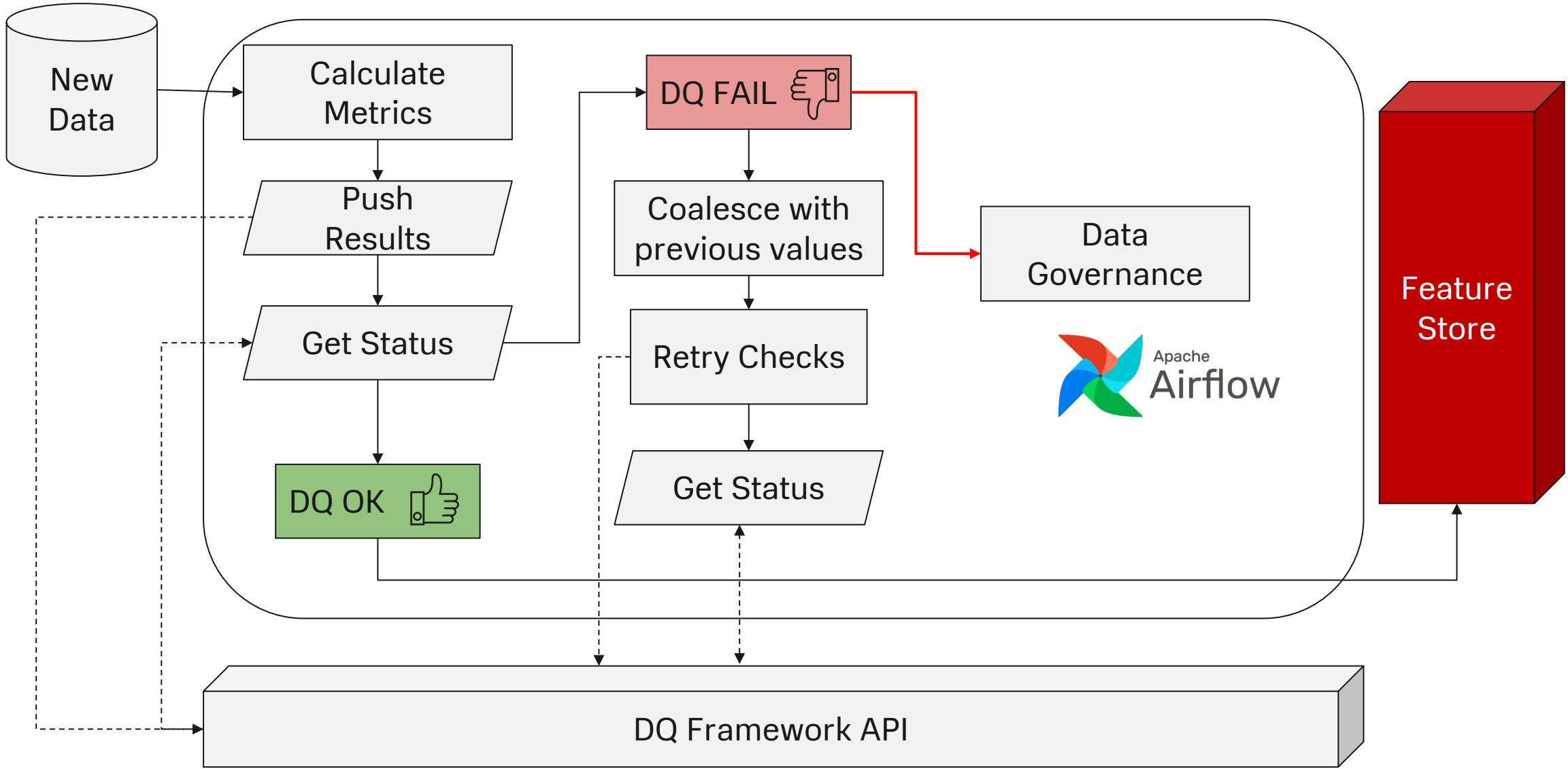
Процессы

Группа признаков в Feature Store



Процессы

Группа признаков в Feature Store



```
graph TD
    NewData[(New Data)] --> CalculateMetrics[Calculate Metrics]
    CalculateMetrics --> PushResults[/Push Results/]
    PushResults --> GetStatus1[/Get Status/]
    GetStatus1 --> DQOK1[DQ OK]
    GetStatus1 --> DQFAIL1[DQ FAIL]
    DQOK1 --> FeatureStore[Feature Store]
    DQFAIL1 --> Coalesce[Coalesce with previous values]
    Coalesce --> RetryChecks[Retry Checks]
    RetryChecks --> GetStatus2[/Get Status/]
    GetStatus2 --> DQOK2[DQ OK]
    GetStatus2 --> DQFAIL2[DQ FAIL]
    DQOK2 --> FeatureStore
    DQFAIL2 --> DataGovernance[Data Governance]
    DataGovernance --> FeatureStore
    DQFrameworkAPI[DQ Framework API] -.-> GetStatus1
    DQFrameworkAPI -.-> GetStatus2
    DQFrameworkAPI -.-> NewData
```


| 3.5

Реализация Инциденты

Инциденты

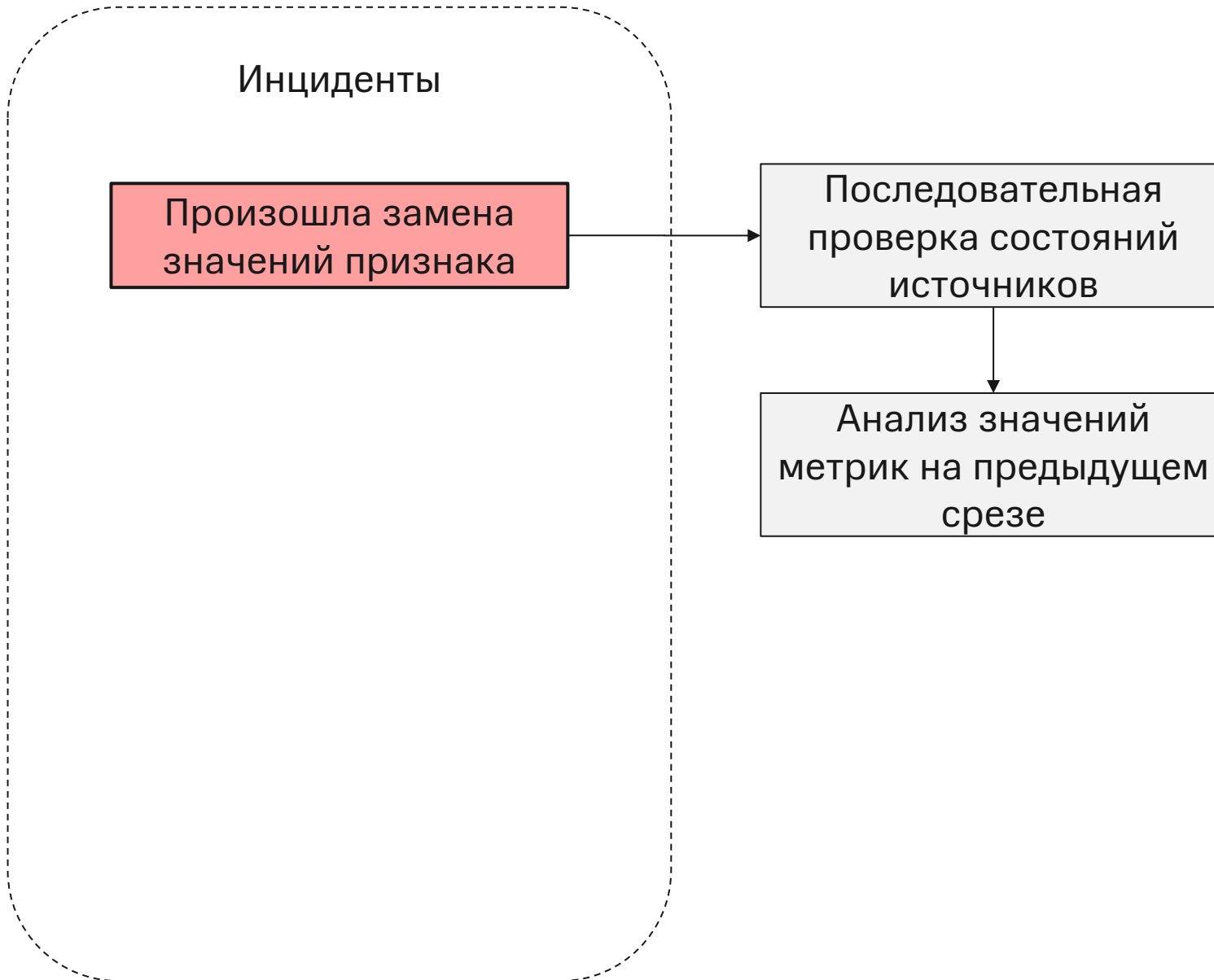
Инциденты

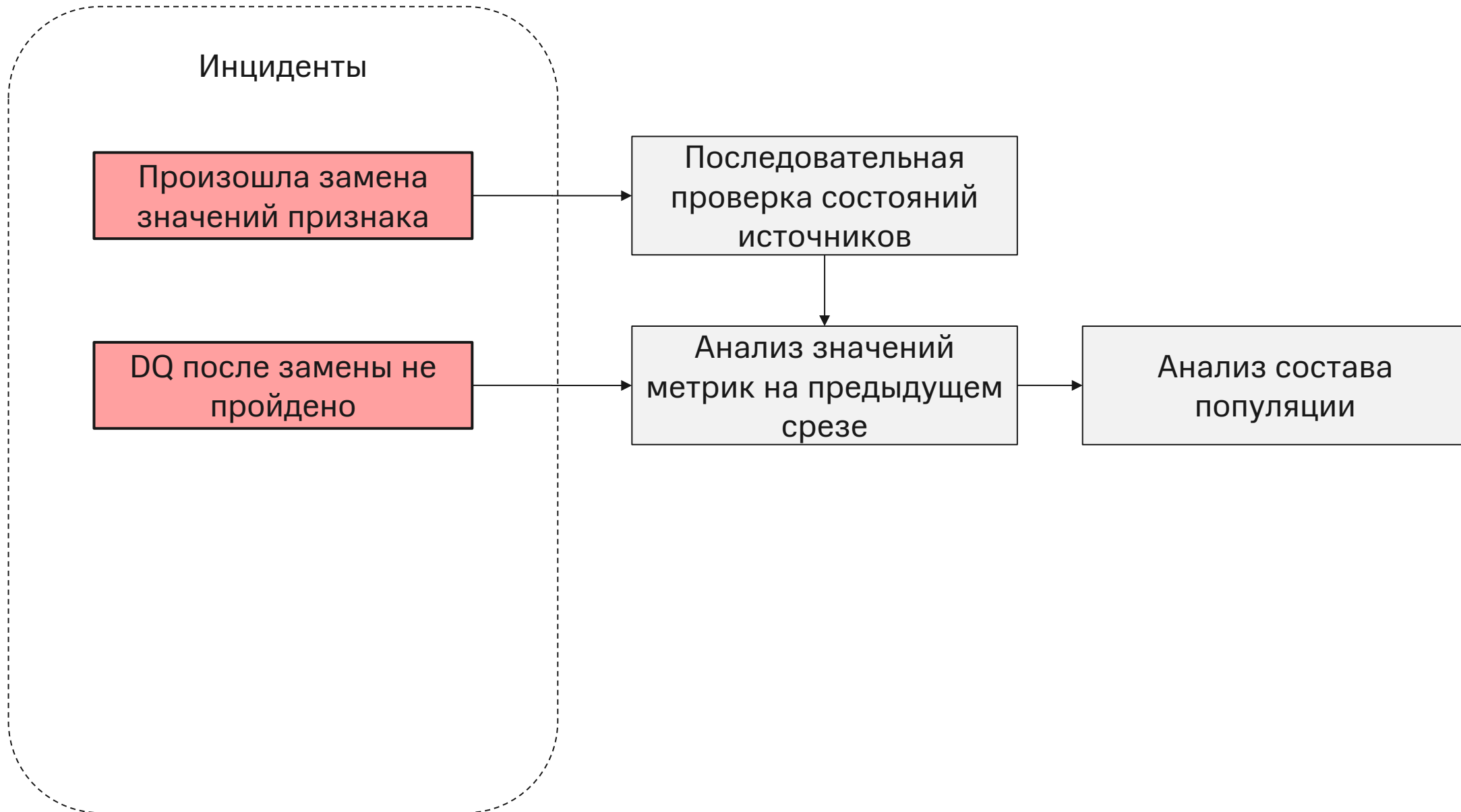
Произошла замена значений признака

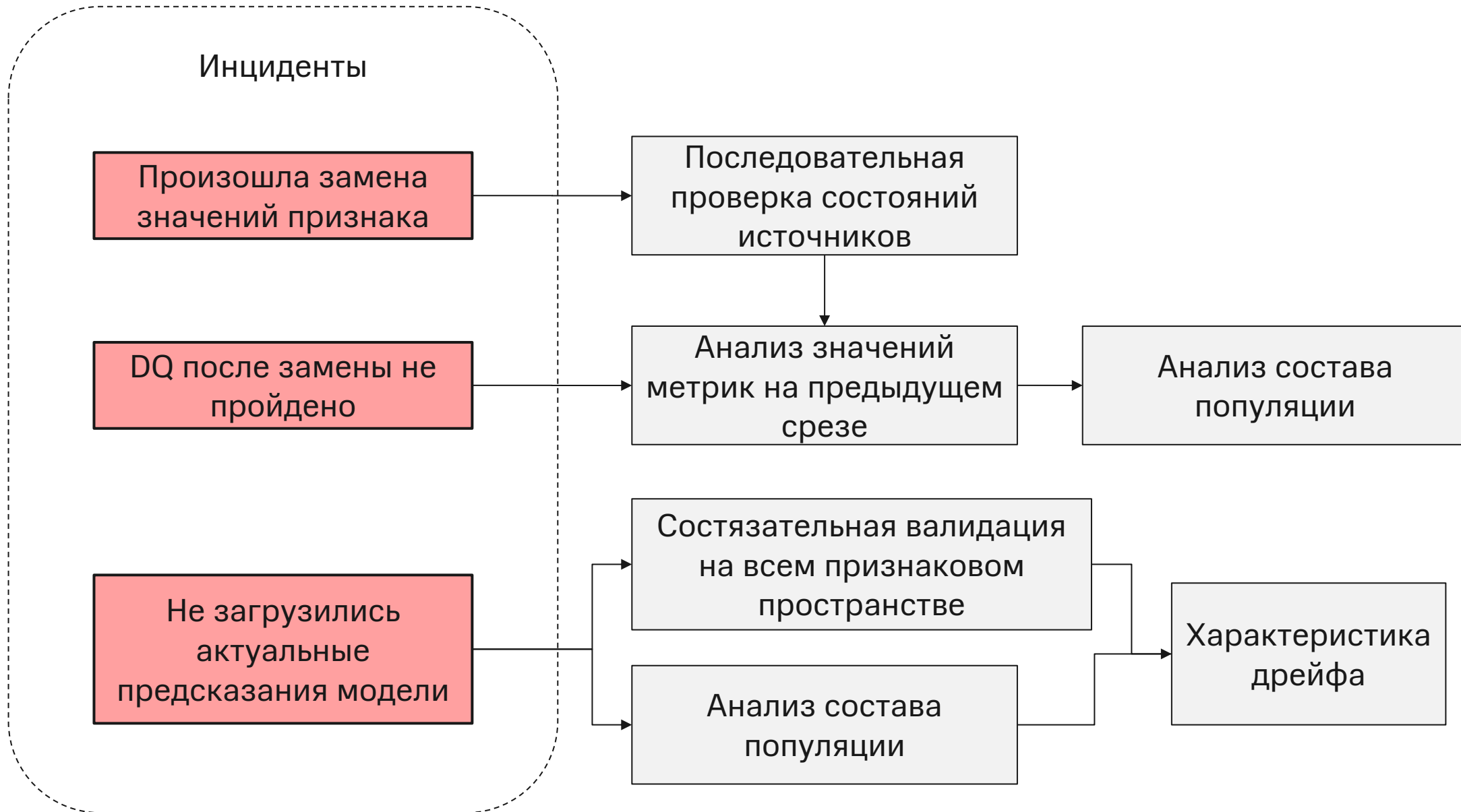
DQ после замены не пройдено

Не загрузились актуальные предсказания модели

Инциденты





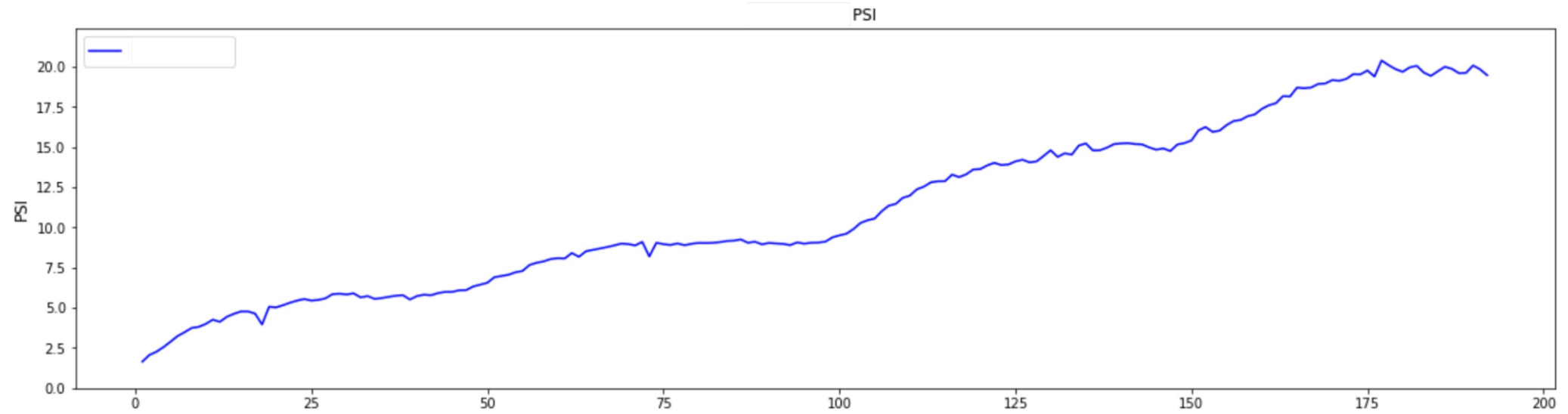


| 4

Примеры

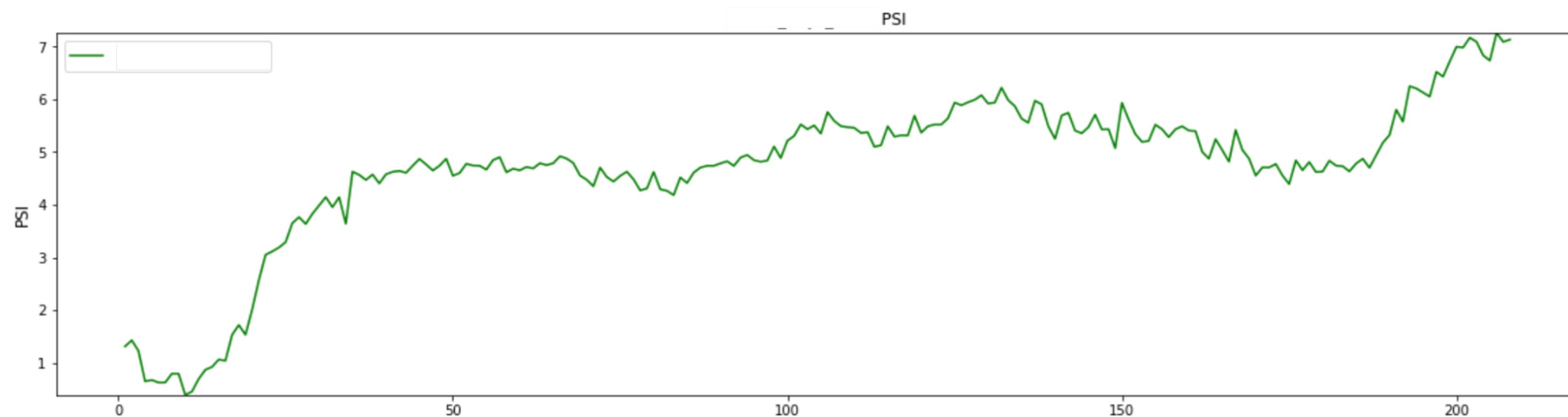
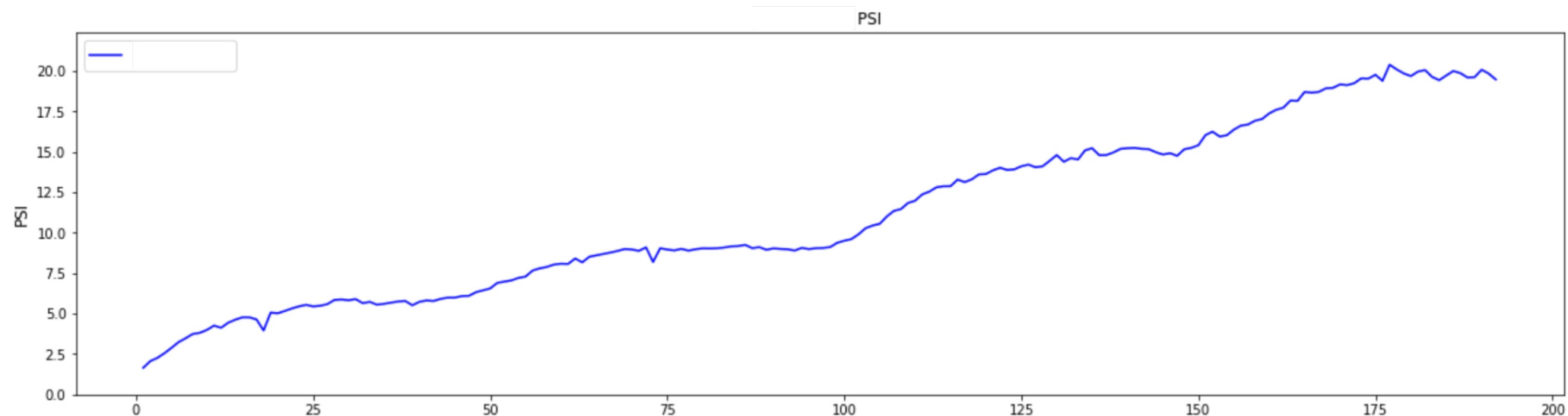
Примеры на исторических данных

Трендовый признак



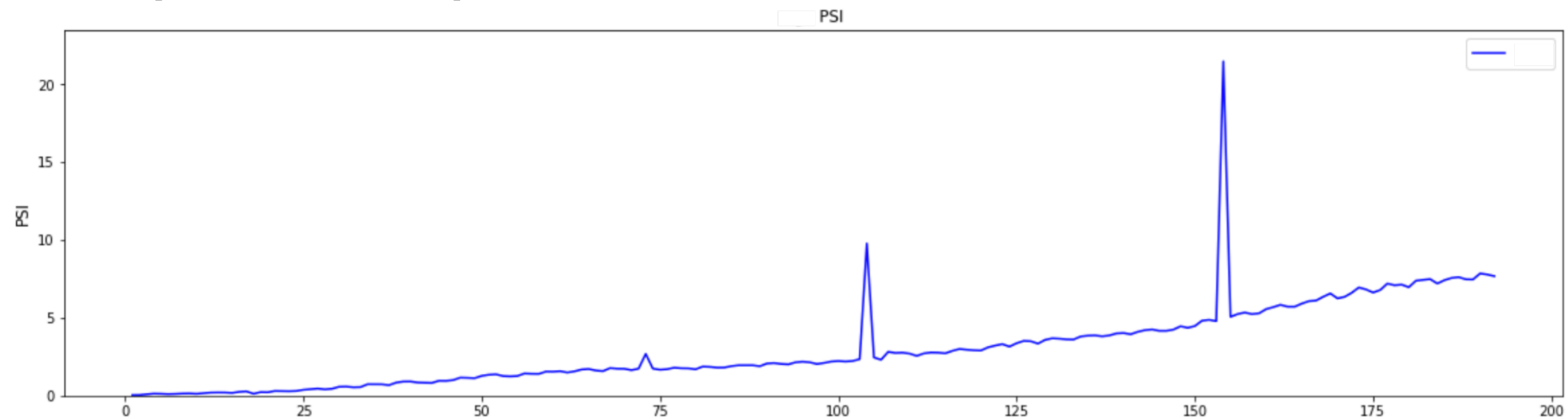
Примеры на исторических данных

Трендовый признак



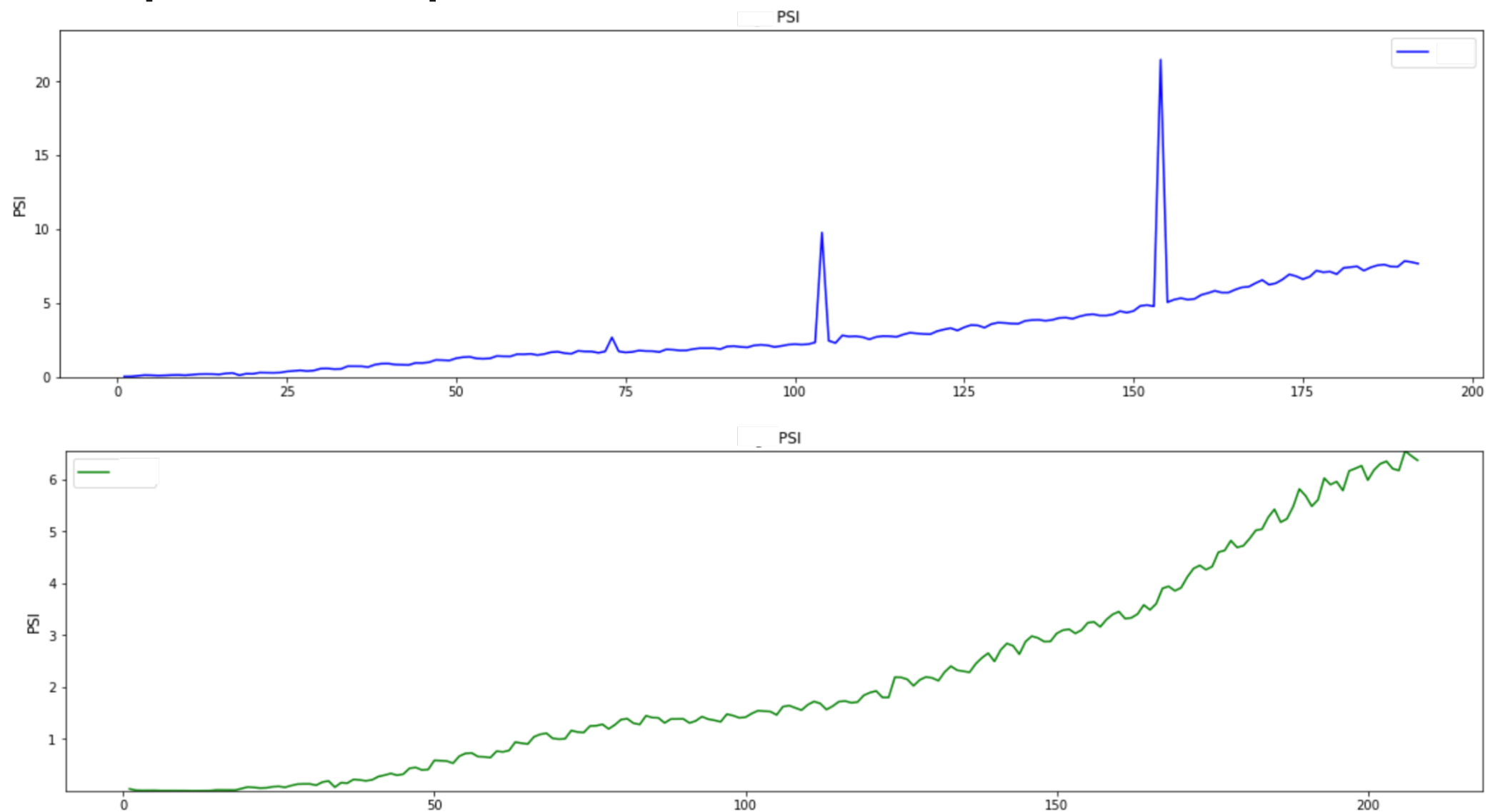
Примеры на исторических данных

Признак с проблемными срезами



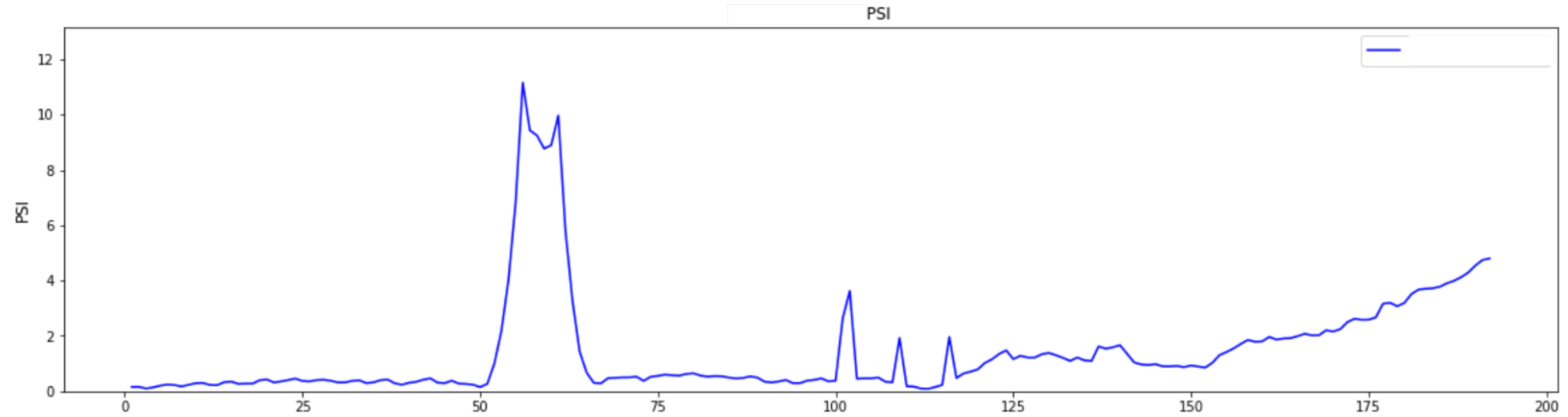
Примеры на исторических данных

Признак с проблемными срезами



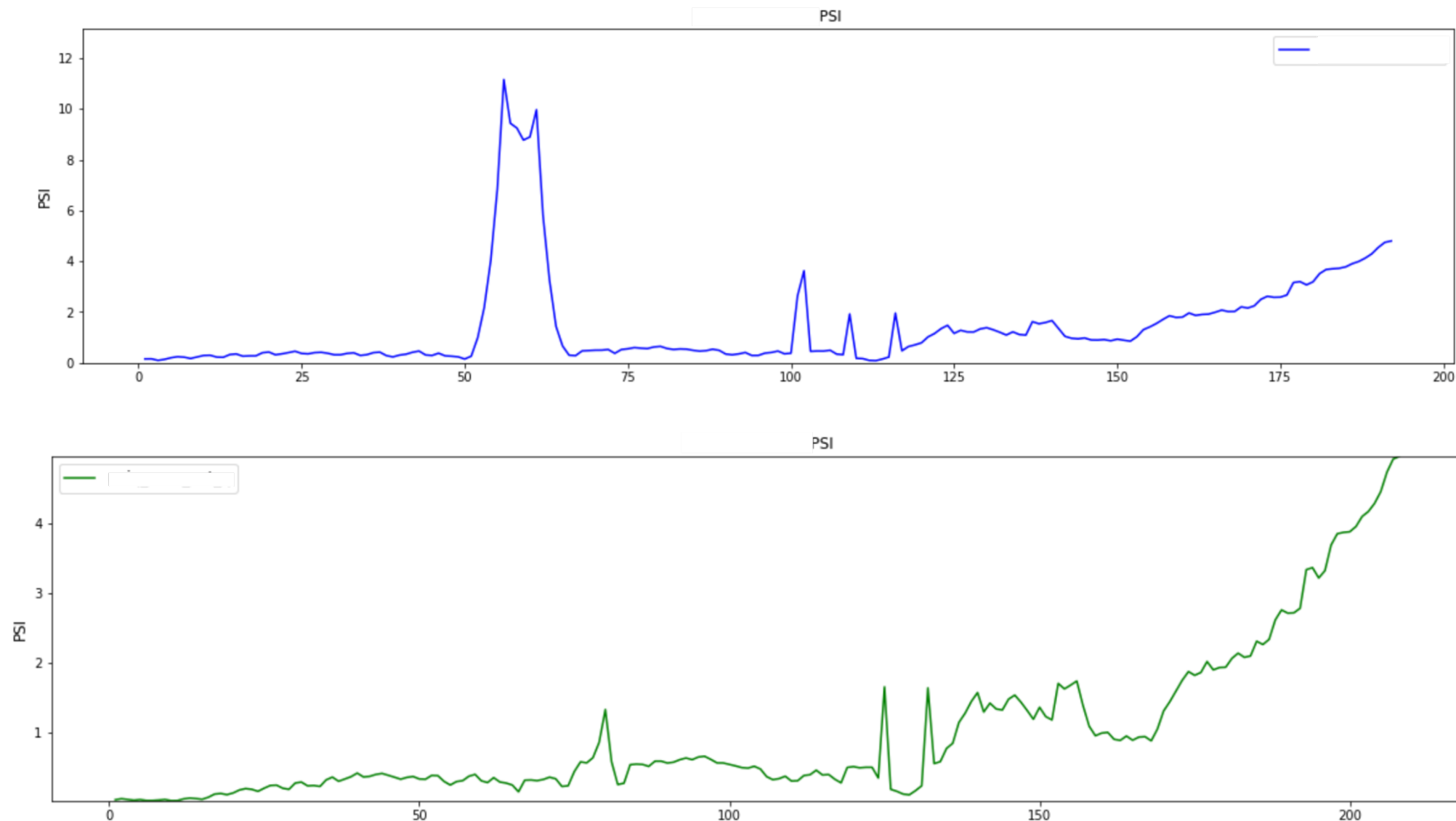
Примеры на исторических данных

Признак с проблемным периодом



Примеры на исторических данных

Признак с проблемным периодом



| 5

Выводы

Выводы

Финальные замечания:

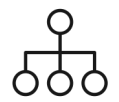


Правильный подход к синтезу стабильных моделей значительно экономит ресурсы при их промышленной эксплуатации

Финальные замечания:



Правильный подход к синтезу стабильных моделей значительно экономит ресурсы при их промышленной эксплуатации

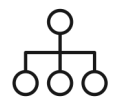


Цель DQ мониторинга – уберечь продуктовые бизнес-процессы от попадания в них ненадежных предсказаний моделей или ошибочных данных

Финальные замечания:

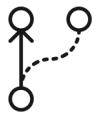


Правильный подход к синтезу стабильных моделей значительно экономит ресурсы при их промышленной эксплуатации



Цель DQ мониторинга – уберечь продуктовые бизнес-процессы от попадания в них ненадежных предсказаний моделей или ошибочных данных

Future Work:

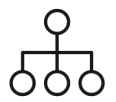


Расширить перечень инцидентов, обрабатываемых автоматически

Финальные замечания:



Правильный подход к синтезу стабильных моделей значительно экономит ресурсы при их промышленной эксплуатации

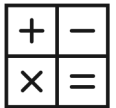


Цель DQ мониторинга – уберечь продуктовые бизнес-процессы от попадания в них ненадежных предсказаний моделей или ошибочных данных

Future Work:



Расширить перечень инцидентов, обрабатываемых автоматически

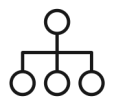


Эффективная реализация многомерных методов детекции дрейфа

Финальные замечания:

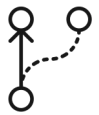


Правильный подход к синтезу стабильных моделей значительно экономит ресурсы при их промышленной эксплуатации

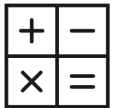


Цель DQ мониторинга – уберечь продуктовые бизнес-процессы от попадания в них ненадежных предсказаний моделей или ошибочных данных

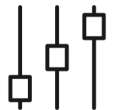
Future Work:



Расширить перечень инцидентов, обрабатываемых автоматически



Эффективная реализация многомерных методов детекции дрейфа



Автоматическая проверка признаков на стабильность при включении в Feature Store

| 6

Q&A