

Рекомендации, поиск и мерчандайзинг хороши ровно настолько, насколько хорош питающий их конвейер. Сделайте четыре вещи правильно: соберите события со всех экранов, осознанно выберите real-time или batch, запустите feature store, дающий одинаковые числа в обучении и проде, и обеспечьте границу приватности данных просмотра. Право о данных просмотра (VPPA) быстро меняется — это инженерное руководство, а не юридический совет; уточняйте детали с юристом.

1 · СОБЕРИТЕ СОБЫТИЯ (сырьё)

- ☐ **Три семейства событий** — playback (play, pause, seek, complete), навигация/clickstream (показ, клик, скролл) и поиск.
- ☐ **Сохраните heartbeat** — пинг плеера «ещё смотрю, на этой позиции» каждые несколько секунд восстанавливает watch time.
- ☐ **Логируйте показы, не только клики** — модель учится на том, что показали и проигнорировали, а не только на запущенном.
- ☐ **Рассчитайте на пик** — 1M одновременно при ~10 событий/мин $\approx 167k$ событий/сек; проектируйте под всплеск, удешевляйте базу.

2 · REAL-TIME ИЛИ BATCH (два темпа)

- ☐ **Stream, когда свежесть меняет ответ** — последний час, тренды, continue-watching; считается за секунды (напр., Flink).
- ☐ **Batch для медленных больших сигналов** — вкус за 90 дней, часы за всё время, эмбединги; по расписанию (Spark), дешевле.
- ☐ **Не тяните real-time везде** — стриминг-инфра дороже и сложнее в эксплуатации; большинству сигналов секунды не нужны.
- ☐ **Lambda (два слоя) против Карра (один)** — лучше один путь кода над логом, если вам реально не нужен отдельный batch-слой.

3 · FEATURE STORE И SKEW (корректность)

- ☐ **Одно определение признака для обоих** — обучение и прод читают одну логику; это и убирает training-serving skew.
- ☐ **Offline + online store** — offline для обучения по всей истории, online для serving с низкой задержкой (~10 мс) в запросе.
- ☐ **Логируйте отданные признаки, учите на них** — если учитесь ровно на отданном, два пути кода не разъедутся.
- ☐ **Point-in-time join** — строка обучения видит только данные до своей метки времени; иначе утечка подделывает точность.

4 · ГРАНИЦА ПРИВАТНОСТИ (данные просмотра)

- ☐ **Не отдавайте «кто-что-смотрел» третьим тегам** — экспозиция по VPPA \$2,500/человек (18 U.S.C. § 2710); ошибка №1.
- ☐ **Конкретное отдельное согласие** — согласие по VPPA не зашито в условия; можно заранее, с понятным способом отозвать.
- ☐ **Срок хранения** — удаляйте PII просмотра в течение года после того, как оно не нужно (VPPA § 2710(e); GDPR).
- ☐ **Псевдонимизируйте и минимизируйте** — работайте по непрозрачному ключу, собирайте нужное, носите согласие с данными.

ПОРЯДОК ДЕЙСТВИЙ — СОБРАТЬ, ПЕРЕДАТЬ, ХРАНИТЬ, ЗАЩИТИТЬ — И КОНВЕЙЕР ЭТО ПРОДУКТ

Всё в discovery — рекомендации, поиск, мерчандайзинг и эксперименты, которые их настраивают, — работает на конвейере данных для персонализации, поэтому именно конвейер, а не модель, чаще всего решает, выигрывает ли стриминговая платформа. Стройте по порядку. Первое: соберите события со всех экранов — playback (heartbeat восстанавливает watch time), навигация/clickstream (показы важны не меньше кликов) и поиск. Второе: передайте их через replayable-лог и выберите темп под каждый признак — stream-обработка за секунды там, где свежесть меняет ответ, batch по расписанию для медленных и больших сигналов; предпочитайте один путь кода (Карра над логом) поддержке одной логики дважды (Lambda). Третье: храните признаки в feature store, который отдаёт одно общее определение и в обучение, и в прод, — это и устраняет training-serving skew, тихий баг, когда признак, посчитанный в обучении одним способом, а на serving другим, незаметно ломает предсказания, — и собирайте обучающие данные через point-in-time join, чтобы строка не видела данных из будущего относительно предсказываемого момента. Четвёртое: обеспечьте границу приватности: данные просмотра юридически чувствительны, поэтому никогда не раскрывайте, кто что смотрел, третьему рекламному или аналитическому тегу (паттерн за волной исков по Video Privacy Protection Act, по \$2,500 на человека), гейтите по конкретному согласию, минимизируйте и псевдонимизируйте, удаляйте по часам хранения. Удобный срез угла — быстрый сторонний пиксель или второй запрос признака, который проще написать, чем переиспользовать, — это и есть дорогой баг или дорогой иск. Относитесь к конвейеру как к продукту. Это инженерное руководство, а не юридический совет — уточняйте детали с юристом.