**Онлайн кинотеатр**

Не будем усложнять кинотеатр, работаем в рамках нашей модели. Считаем, что пользователь выставляет рейтинг фильму после его просмотра, хотя, как правило, пользователь (зарегистрированный) может давать оценку фильмам, не посмотрев их у нас. Также должны быть данные о посещаемости онлайн-кинотеатра, т.е. сколько пользователей заходили на сайт, просматривали анонсы, трейлеры каких то фильмов.

1. **Основные бизнес - отчеты**

* Динамика количества просмотров (в нашем случае эквивалентно количеству платежей)
* Количество пользователей, в т.ч. активных, т.е. смотрящих в среднем за месяц более N фильмов.
* Конверсия посетителей сайта в зарегистрированных пользователей онлайн-кинотеатра.

1. **Основные данные и источники их поступления**

* Пользователи. Для просмотра фильмов необходимо зарегистрироваться, указав: возраст, пол, профессию, номер телефона. Автоматически присваивается уникальный номер *user id* и дата регистрации*.*
* Фильмы. За наполнение фильмотеки отвечает выделенный сотрудник (или отдел), назовем его “Менеджер по фильмотеке”. При поступлении нового фильма в фильмотеку указываются: название, дата релиза, URL-ссылка, категория. Фильм может быть отнесен к нескольким категориям. Автоматически присваивается уникальный номер *movie id* и дата регистрации*.*
* Данные о просмотрах. При просмотре фильма фиксируются: код пользователя *user id,* код фильма *movie id,* дата просмотра, рейтинг (пользователь выставляет после просмотра).

1. **Основные сущности и процесс заливки данных**

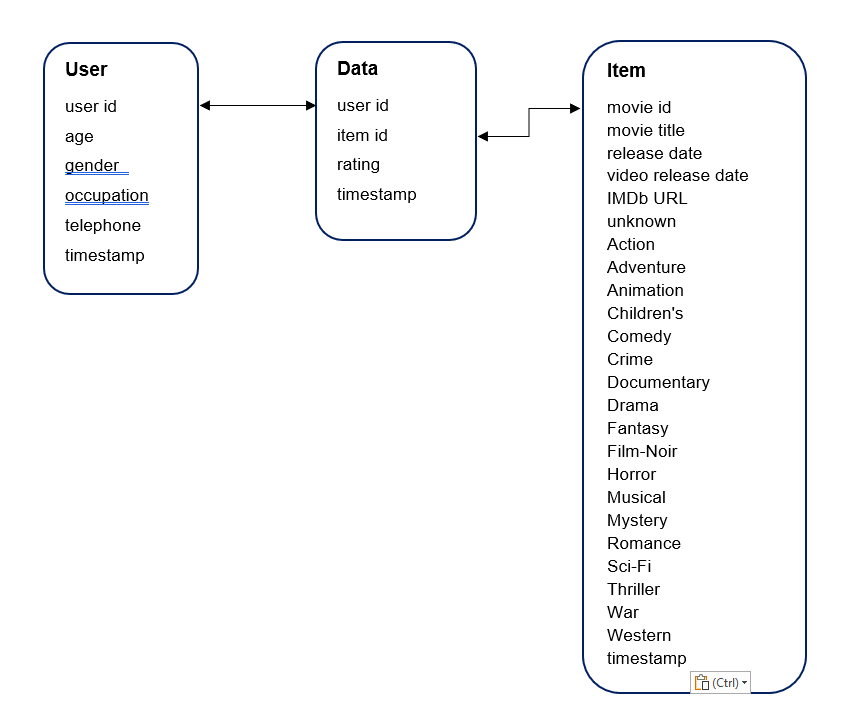


Таблица User. Данные заносит пользователь при регистрации в Онлайн-кинотеатре.

Таблица Item. Данные заносит менеджер продукта при добавлении нового фильма в фильмотеку.

Таблица Data. Данные user id, item id, timestamp заносятся автоматически при выборе и просмотре фильма, rating заносит пользователь после просмотра фильма.

1. **Основные проверки на качество данных**

* Поле age таблицы User. Проверка при вводе: числовое значение в интервале от 10 до 100. Поле обязательное для заполнения.
* Поле gender таблицы User. Выбор из списка (F, M) при вводе. Поле обязательное для заполнения.
* Поле occupation таблицы User. Выбор из списка при вводе. Поле обязательное для заполнения.
* Поле occupation таблицы User. При большом количестве пользователей, указавших профессию “other”, возможно корректировки на основе информации полученной, например, при анкетировании, опросе пользователей сайта. Расширение предлагаемого списка профессий.
* Поле telephone таблицы User. Поле обязательное для заполнения. При регистрации пользователю отправляется смс с кодом для подтверждения валидности указанного номера.
* Поле rating таблицы Data. Выбор из списка возможных значений от 1 до 5 (или 1-10).
* Поле movie title таблицы items. Автоматическая проверка при вводе на наличие фильма с таким названием в фильмотеке.
* Возможны ошибки при вводе названия фильма, особенно с названиями на иностранном языке. Периодические проверки наличия фильмов с похожими названиями. Корректировка, если это идентичные объекты.
* Поле IMDb URL таблицы items. Автоматическая проверка при вводе на наличие фильма с такой ссылкой в фильмотеке.
* Жанр фильма таблицы items. Выбор из списка, обязательное заполнение при вводе нового фильма. Допускается выбор нескольких значений.
* Периодические проверки фильмов, отнесенных к нескольким жанрам, на предмет совместимости указанных жанров.

1. **Data-проект, который должен улучшить показатели бизнеса онлайн - кинотеатра.**

Владелец бизнеса недоволен финансовыми показателями работы онлайн - кинотеатра. Одной из причин он считает не отвечающий запросам пользователей состав фильмотеки, как то: популярность фильмов, их количественный состав по жанрам. Подтверждением этому он считает:

* низкий уровень конвертации посетителей сайта в зарегистрированных пользователей кинотеатра
* большая доля пользователей с низкой активностью просмотров
* большая доля пользователей в узком возрастном диапазоне
* ряд жанров, имеющих более высокий средний рейтинг, количественно значительно уступают жанрам, имеющим более низкий рейтинг

Поставлена задача построить модель, дающую рекомендации по формированию фильмотеки.

Функционально компанию представляют: владелец продукта, менеджер продукта, аналитик данных и специалист по БД - разработчик.

**5.1 Business understanding**

На данном этапе необходимо четко сформулировать задачу бизнеса, а именно:

* Метрики, их количественные показатели, описывающие текущее состояние проблемного участка бизнеса
* Целевые показатели выполнения бизнес-проекта, критерии оценки. Допустимые интервалы значений целевых показателей, при которых результат будет считаться эффективным.
* Финансовые показатели, которые планирует получить собственник бизнеса при успешном выполнении проекта.
* Объем допустимых затрат в рамках проекта, какие могут быть выделены ресурсы.
* Что будет результатом выполнения проекта: отчет, рекомендации владельцу продукта, дополнительный сервис и т.д.
* Возможные риски, которые могут повлиять на реализацию проекта
* Создание плана проекта

Участники: владелец бизнеса, владелец продукта, аналитик.

**5.2.** **Data understanding**

* Описание и анализ имеющихся в распоряжении данных.
* Методы и инструменты, которые будут использоваться в рамках проекта.
* Определение ключевых аттрибутов, их наличие в имеющихся данных
* Количественная достаточность данных для моделирования, проверки гипотез и расчета статистических показателей
* Качество данных, необходимые действия для их подготовки

Участники: аналитик

**5.3 Preparation**

* Выполнить очистку данных (исправление ошибок, заполнение пропусков и т.д.)
* Сформировать датафреймы, необходимые для проведения анализа и моделирования.
* Выбрать признаки, разделить данные на обучающие и тестовые выборки
* Задокументировать этапы подготовки данных

Участники: аналитик, специалист по БД - разработчик.

**5.4 Modeling**

* Выбрать алгоритмы и провести моделирование.
* Оценить результаты и определить, являются ли они понятными и достаточными для формирования возможных решений задачи проекта или требуются дополнительные исследования.

Участники: аналитик.

**5.5 Evaluation**

* Оценка результатов.
* Выполнена ли задача полностью в рамках целевых показателей, какие дальнейшие шаги по внедрению результатов проекта
* Если представленное решение не отвечает в полной мере установленным требованиям, можно ли использовать результат в какой то части, какова эффективность этих решений.
* Возможно появление неких инсайтов в ходе исследований, выявление параметров, влияющих на бизнес, которые не рассматривались на этапе планирования проекта. Какие по ним решения.

Участники: владелец бизнеса, владелец продукта, аналитик.

**5.6 Deployment**

* Составить план внедрениярезультатов моделирования, включая дальнейшую реализацию, обслуживание и мониторинг эффективности.
* Провести итоговый обзор проекта

Участники: владелец продукта, аналитик, специалист по БД - разработчик.

1. **Требуемые роли в команде по работе с данными на этапах 4 и 5**

Указаны непосредственно в п. 5. На этапе 4 менеджер продукта отвечает за наполнение данных в фильмотеке, проверку качества и корректности данных выполняет специалист по БД - разработчик, привлекая, где необходимо, к исправлению ошибок менеджера продукта.